Université de Montréal

Feature selection and term weighting beyond word frequency for calls for tenders documents

Par

Qing Ma

Département d'informatique et de recherche opérationnelle

Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
En vue de l'obtention du grade de Maître ès science (M.Sc.)
en informatique

Décembre, 2006

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :

Feature selection and term weighting beyond word frequency for calls for tenders documents

présenté par :

Qing Ma

a été évalué par un jury composé des personnes suivantes :

Esma Aïmeur, président-rapporteur
Jian-Yun Nie, directeur de recherche
Philippe Langlais, membre du jury

Mémoire accepté le 20 avril 2007

RÉSUMÉ

La classification de textes est un processus qui assigne automatiquement des documents des textes aux catégories prédéfinies. Afin de classifier des textes, nous devons extraire de bonnes caractéristiques à partir d'eux. Les méthodes telles que la sélection de caractéristiques et la pondération qui permettent de distinguer les bonnes caractéristiques et des mauvaises, ont été développés pour améliorer le résultat de classification.

Dans notre travail, nous étudions le problème de la classification des documents d'appel d'offre. À la différence du document typique(les nouvelles), les documents de l'appel d'offre contiennent un bon nombre d'information procédurale indépendant du sujet des documents. Information de position, entités nommées et concepts sont les trois facteurs qui peuvent distinguer de bonnes caractéristiques des mauvaises dans les documents d'appel d'offre. Dans ce mémoire, nous proposons les méthodes de sélection de caractéristiques et les méthodes de pondération de poid qui comptent l'information de position, les entités nommées et les concepts pour mesurer l'importance de la phrase.

Pour vérifier l'effet des deux méthodes proposées, nous avons entrepris des expérimentations en utilisant les classificateurs Naïve Bayes et SVM sur les documents de l'appel d'offre de type FBO. Nous obtenons le meilleur résultat quand nous employons les méthodes de pondération qui combinent tous ces facteurs pour les classificateurs Naïve Bayes et SVM. On observe une légère amélioration sur les resultants de classification après des expériences avec la methode de sélection de caractéristiques. Nous observons également que la méthode sélection de caractéristiques par le filtrage de

phrase et la méthode de pondération améliorent le classificateur Naive Bayes par une plus grande marge que le classificateur SVM.

Mots-clés: classification, pondération, sélection de caractéristiques, appel d'offre

ABSTRACT

Text classification is a process that automatically assigns text documents to predefined categories. In order to classify text documents, we must extract good features from them. Methods such as feature selection and term weighting, that allow to distinguish good feature and bad features in a document, have been developed to improve the classification result.

In our study, we investigate the problem of classification of call for tender (CFT) documents. Unlike the typical news document, CFT documents contain lots of procedural information unrelated to the subject of the documents. Sentence position information, named entities and concepts are factor that can distinguish good features and bad features in the CFT documents. In this paper, we propose feature selection methods and term weighting methods that rely on position information, named entities and concepts to measure importance of sentence.

To verify the effect of both proposed methods, we conducted experiments using Naïve Bayes and SVM classifiers on FBO call for tender documents. We obtain the best result when we use term weighting methods that combine all factors for the Naïve Bayes and SVM classification. Only slight improvement is observed after we conduct feature selection experiments. We also observe that for both feature selection by sentence filtering and term weighting method improves the performance of Naive Bayes classifier by a bigger margin than the performance of SVM classifier.

Keyword: classification, term weighting, feature selection, call for tender

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I'd like to thank my thesis supervisor, Jian-yun Nie. Thank you for everythng you've done for me. You are an amazing scientist. Your encouragement is a constant inspiration for me. Without you, my work wouldn't be possible.

Also, I'd like to thank all the people from RALI group who has helped me and supported me throughout my research: Francois Paradis, Shi Lixin, Graham Russell, etc. Thank you for helping me with my research whenever I had problems. Words cannot express my gratitude.

Special thanks to my family, which I value most in my life; you are my source of strength, love, happiness, and support. My father, Hewu Ma, my mother Shijin Yang, are always behind me good times and bad times. Thank you for listening to me and making me believe that nothing is impossible. I dedicate this dissertation to all of you.

Qing Ma

January 2007

CHAPTER 1

INTRODUCTION

Text classification is a very dynamic area of research in artificial intelligence. The purpose of text classification is to classify text documents into a pre-defined set of categories. There are many algorithms, both supervised and unsupervised, such as SVM (Joachims, 1998), Naïve Bayes (McGallum et al., 1998), KNN (Yang et al, 2002), that can be used to accomplish this task.

In this thesis, we are interested in the classification of call for tender (CFT) documents. It's a type of document in which an authority (or solicitor) specifies his/her requirement for goods or services so that a contractor can submit a tender. As e-commerce develops rapidly, more and more industries are interested in finding out business opportunities on-line. For example, *fedbizopps.gov* or *merx.ca* are two large websites where government calls for tenders are published. Online CFT is an important source of business opportunities. However, most CFT are not well organized. It is difficult for users to locate the ones that are relevant to them. Therefore, automatic classification of the CFT document is required in order to organize automatically calls for tenders into a class hierarchy, so that users can browse through.

In previous studies on automatic classification, it is found that the performance of classifiers usually depends on two important aspects. The first is the classification algorithm and the second is the relevance of the features used in the algorithms. As there are many classification algorithms which can be used for the purpose of CFT classification, we will focus on the second problem in this thesis. In previous studies, features are usually words (except stopwords) extracted from documents. In many cases, the frequency of occurrences of words is also used to weight the importance of them. This strategy works well for the classification of general documents, for example, newspaper articles. In some other cases, it is found that word extraction according to their frequencies alone is not sufficient: some of the words represent important meanings of the documents, while some others do not. These latter are considered to be irrelevant features or noise. However, there have not been many studies focusing on the extraction of relevant words or features for specific documents such as CFT. In the case of CFT, it is especially crucial to extract relevant features and filter out irrelevant ones, because the important meanings (the subject of the calls, i.e. the goods or services required) are not described by repetitive words, thus by words with high frequency. Instead, the subject is usually described by one or a few sentences. On the other hand, there is a large portion in CFT documents specifying the procedure for a submission. This part of the document is not directly related to the subject of the call, and is not useful for CFT classification. We call this part "procedure noise". To illustrate the problem of "procedural noise", here is an example of call for tender:

"Landfill Disposal Services. Landfill Disposal Services for solid waste generated and delivered by various Navy installations within Navy Region

Northwest. Contractor shall accept and dispose of solid waste delivered to facility. Commodities types include but not be limited to, municipal solid waste, creosote pilings, soil, concrete, asphalt, asbestos, treated lumber, asphalt & concrete with rebar, construction demolition and land clearing debris, sludge cake/filter press dewatered solids, petroleum contaminated soils and truck wastes. *Landfill service provider shall ensure compliance with all federal, state, and local laws or regulations related to refuse disposal and can be licensed by Washington State Utility and Transportation Commission. Electronic monthly billing services to include record of Daily transaction detail by date, time, net weight in tons, commodity delivered and vehicle identification number. Estimated monthly delivery of waste: 2500 Tons. This procurement will be classified under XXX Code (North American industry classification system): XXX. Size standard for this code is 10.5M.*

This acquisition is being solicited using commercial procedures in accordance with FAR Part 12. Firm Fixed Price Contract for the period of 27 July through 30 September 2002. To be eligible for award, prospective offerors must be registered in Central Contractor Registration (CCR) Internet site: http://www.ccr.gov. Link to FedBizOpps document. "

The underlined part is related to the subject of the call. The text in italics is the standard procedural submission information present in many CFT documents. Those types of information are irrelevant to the subject of the call for tender document, although

they are useful for users to determine later if they are eligible. For CFT classification, our purpose is to organize all the CFTs according to the subject of the calls. Therefore, the procedural part of CFT should be filtered out.

The problem of irrelevant or noise features also occurs in traditional classification. Two main techniques have been used to select good features from a set of candidate ones: feature selection and term weighting. Both feature selection and term weighting methods try to distinguish relevant and irrelevant features in the document. Traditionally, the feature selection methods involve eliminating irrelevant contents by using statistical method such as information gain (InfoGain) or document frequency (DF) threshold. For term weighting, traditionally one uses the frequency of words in the document: the more a word occurs in a document, the more it is considered to be important and related to the subject. However, previous studies show that it's difficult to eliminate the irrelevant features in the call for tender documents effectively by using those standard methods (Paradis, 2005) due to the very fact that there is only a concise description of the subject, while a lengthy description of the submission procedure is usually included.

The goal of this dissertation is to identify better factors to help determine relevant/irrelevant features in the call for tender documents, so that the classification performance can be improved. In this study, we examine three types of feature: position of the sentence, named entities contained in the sentence and concepts in the sentences. These factors are strongly related to the general characteristics of CFT. In general, important sentences that describe the subject appear at the beginning of a CFT. Therefore, the position of the sentence is an indicator of relevance or irrelevance of

sentence. Named entities (NE) such as organization, place, email address, date, money, etc., are frequently used in CFT. Some named entities are helpful in suggesting relevance of a sentence, while some other can suggest irrelevance. As an example, the contact URL address of the authority often appears in the procedural part of CFT; thus a sentence containing such a NE is usually irrelevant to the subject. For example, in the CFT document shown above, the sentence "*To be eligible for award, prospective offerors must be registered in Central Contractor Registration (CCR) Internet site: http://www.ccr.gov. Link to FedBizOpps document.* ", which contains a URL address, is an irrelevant sentence. Therefore URL can be an indicator of irrelevance of the sentence.

The description of the subject of the call often use terms that correspond to specialized concepts. For example, "Landfill Disposal Services", "treated lumber" in the above example correspond to a concept. The occurrence of specialized concepts in a sentence is also a useful indicator on the relevancy of the sentence to the subject.

Therefore, in this dissertation, I will exploit mainly position, named entities and concepts or combination of all these factors to distinguish relevant and irrelevant sentences to the subject, and to assign better weights to terms (or features). We hope that with this refined feature selection and weighting, the classification result of the CFT documents can be improved compared to the classification using traditional feature selection and term weighting methods.

We will evaluate and compare feature selection and term weighting method by conducting experiments using SVM and Naïve Bayes classifiers. Specifically, we seek the answers to the following questions empirically:

1. Can feature selection or term weighting method based on these additional factors lead to better performance with commonly used classification algorithms such as Naïve Bayes and SVM for CFT documents? How do they compare to traditional feature selections such as the approach based on information gain?
2. Which method can achieve better classification performance for Naïve Bayes and SVM classifiers, feature selection or term weighting method?
3. Which factors among concepts, location information and named entities seem to have the most impact to improve the performance of SVM and Naïve Bayes classifiers?

This study has been carried out within a research project in collaboration with Nstein Technologies, entitled "Matching Business Opportunities on the Internet" (MBOI). It aims to automatically identify business opportunities on the Internet for small businesses. The whole research for MBOI project covers the following aspects: automatically collecting calls for tenders on the Web, matching calls for tenders and the need of an enterprise, classifying calls for tenders, searching calls for tenders in different languages. The study described in this thesis is only concerned with the aspect of classification of call for tender documents.

Although our focus is on a specific type of document – CFT, we argue that the general idea to exploit specific characteristics of documents should also be used on other types of documents. Therefore, this study can be seen as a case study of this general idea. We will show in our experiments that by incorporating specific factors to determine the

relevance/irrelevance of sentences, and to weigh terms accordingly, we can achieve better classification performance.

The rest of this thesis is organized as follows. in Chapter 2, I will first introduce the classical classification algorithms - SVM and Naïve Bayes, as well as the basic feature selection and term weighting method for classification. In chapter 3, I will describe the three relevance distinguishing factors studied in this thesis. These factors will also be related to some previous studies on feature selection and term weighting method using similar factors. Chapter 4 will present the feature selection and term weighting methods I use in this thesis to improve the classification and I will discuss them in detail. In Chapter 5, we discuss the experiments carried out with those methods and the empirical results in our experiments. The final section presents some conclusions and recommendation for future research.

CHAPTER 2

CLASSIFICATION, FEATURE SELECTION AND TERM WEIGHTING – AN OVERVIEW

We are all faced with an overflow of information in our everyday life. The development of different medium such as Internet and TV contributes to this phenomenon. The question is how we can transform information into something useful for us. Information needs to be extracted, analyzed and organized before it is delivered to us. Information retrieval (or search engine) is a useful tool for this purpose: It allows the users to find documents related to a query describing his/her information requirement. However, it is also know that users often cannot describe their information need with the appropriate words. In some cases, a better solution is to offer the users with a well organized document structure, so that they can browse. The structure can correspond to the semantic categories of documents, which are defined manually by domain experts. The purpose of document classification is to assign appropriate categories to documents.

Manual classification has been used for a long time before computer is popularized. However, the classification by human is a very slow and tedious process. The amount of pages needed to be indexed grows much faster than human indexers can process. Therefore, it's important to classify information by using automatic classification techniques that learn from human classification examples, and is able to

classify new documents based on what it has learnt. In this chapter, we will review the general problem of automatic classification and the basic methods used for it.

2.1 Overview of Automatic Classification

The automatic classification is a procedure that consists of constructing a model that tries to correctly predict class of different objects. The model is built based on a set of variables describing different characteristics of the objects (i.e., independent variables) and on a training set of previously labeled item. In text classification, words are usually used as such variables (in some cases, feature extraction is used, which tries to generate more abstract features from words and documents). Once the model is built, it can be used to classify new objects whose class information previously unknown.

General classification methods can be applied to many different areas such as texts and bio-information. In text classification, we try to assign a document to one or more categories (classes), based on its contents. Document classification can be used to deal with problems such as spam filtering and web pages classification.

2.2 Text Classification Methods

Text classification methods can be divided into two categories: supervised classification where some external mechanism (such as human feedback) provides information on the correct classification for documents, and unsupervised document classification, where the classification must be done entirely automatically without reference to external information. In this thesis we only deal with supervised document classfication. We assume that there is a set of manually labeled texts that we can use as our training data.

Many different algorithms have been developed for document classification over the years. The most commonly used ones include Naïve Bayes Algorithm, k-Nearest Neighbor (KNN), Support Vector Machine (SVM), Rocchio, etc. For our study we only used Naive Bayes and SVM classifier. Both classifiers are known to perform well in text classification and are good choices as classification baseline. However, Naïve Bayes is known to be sensitive to feature selection and term weighting but SVM is not. This fact also helps us to know better about the effect of our proposed term weighting and feature selection method.

2.1.1 Naive Bayes Classification

Naïve Bayes (NB) algorithm has been widely used for document classification, and shown to be fast and produce very good performance (Mitchell, 1996). The basic idea is to use the joint probabilities of words (used as features) and categories to estimate the probabilities of categories given a document. NB algorithm computes the posterior probability that the document belongs to different classes and assigns it to the class with the highest posterior probability. Bayes Theorem is the basis of the NB algorithm. The posterior probability of class is computed using Bayes rule and the testing sample is assigned to the class with the highest posterior probability. In the following, we will provide more details of this algorithm.

2.2.1.1 Bayes Theorem

Bayes's theorem is a result in probability theory. It relates the conditional and marginal probability distribution of random variables. As we know, the probability of an event A conditional on another event B is generally different from the probability of B conditional on A. However, there is a definite relationship between the two, and Bayes's theorem is the statement of that relationship.

Let B be the data record (case) whose class label is unknown. Let A be some hypothesis, such as "data record B belongs to a specified class C." For classification, we want to determine P(A|B) -- the probability that the hypothesis A holds, given the observed data record B.

P(A|B) is the posterior probability of A conditioned on B. For example, if we use A to express the statement that a fruit is an apple, and B the statement that it is red and round, then P(A|B) expresses the probability that a fruit is an apple, given the condition that it is red and round. In contrast, P(A) is the prior probability, or a priori probability, of A. In this example P(A) is the probability that any given data record is an apple, regardless of how the data record looks like. The posterior probability, P(A|B), is based on more information (such as background knowledge) than the prior probability, P(A), which is independent of B.

Similarly, P(B|A) is posterior probability of B conditioned on A. That is, it is the probability that B is red and round given that we know that it is true that B is an apple. P(B) is the prior probability of B, i.e., it is the probability that a data record from our set of fruits is red and round. Bayes theorem is useful in that it provides a way of calculating

the posterior probability, P(A|B), from P(A), P(B), and P(B|A). Bayes theorem is as follows:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

In the next section, we will see how Bayesian theorem can be used to derive the Naïve Bayes probabilistic model.

2.2.1.2 Naïve Bayes probabilistic model

The probability model for a classifier is a conditional model which is represented by the following formula:

$$P(C \mid Wi....Wn)$$

over a dependent class variable C with a small number of outcomes or *classes*, conditional on several feature variables W_1 through W_n

Using Bayes Theorem, we can write:

$$P(C \mid W_1 ..., W_n) = \frac{P(C)P(W_1,..., W_n \mid C)}{P(W_1,..., W_n)}$$

Since the denominator is a constant and doesn't depend on the C, we can derive the following formula:

$$P(C, W_1 ..., W_n) \propto P(C)P(W_1,..., W_n \mid C)$$

$$= P(C)P(W_1 \mid C)P(W_2,...W_n \mid C, W_1)$$

$$= P(C)P(W_1 \mid C)P(W_2 \mid C, W_1)P(W_3, ...W_n \mid C, W_1, W_2)$$

We assume that each feature W_i is independent of every other feature W_j for i ≠ j, the joint probability can be expressed as:

$$P(C, W_1 ..., W_n) = P(C)\prod_{i=1}^{n} P(W_i \mid C)$$

The conditional distribution over the class variable C can be expressed as follows:

$$P(C \mid W_i ....W_n) = \frac{1}{Z} P(C)\prod_{i=1}^{n} P(W_i \mid C) \propto P(C)\prod_{i=1}^{n} P(W_i \mid C)$$

where Z is a scaling factor dependent on features $W_i ....W_n$.

2.2.1.3 Naïve Bayes Classifier

The Naïve Bayes classifier can be derived from the Naïve Bayes probability model.

The naive Bayes classifier combines this model with a decision rule. One common rule is to pick the hypothesis that is the most probable; this is known as the maximum posteriori (MAP) decision rule. Assume that a document is represented by a set of words $W_1 ...W_n$ appearing in it. These words are called features. The corresponding classifier is the classification function $C(W_1 ...W_n)$ defined as follows:

$$C(W_1 ....W_n) = \text{argmax } P(C = c) \prod_{i=1}^{n} P(W_i = w_i \mid C = c)$$

Naïve Bayes assumes that all attributes of examples are independent of each other given the context of the class. While this assumption is clearly false, in most real-world tasks, Naive Bayes often performs very well in classification. This paradox is explained by the fact that classification estimation is only a function of the sign (in binary cases) of the function estimation; the function approximation can still be poor while classification accuracy remains high (Friedman 1997; Domingos and Pazzani 1997). Because of the independence assumption, the parameters for each attribute can be learned separately, and this greatly simplifies learning, especially when the number of attributes (W_i) is large. Document classification is just such a domain with a large number of attributes. The attributes of the examples to be classified are words, and the number of different words can be quite large indeed – usually in the order of hundreds of thousands.

Studies comparing classification algorithms have found the Naïve Bayes classifier to be comparable in performance to classification trees and to neural network classifiers. They have also exhibited high accuracy and speed when applied to large databases.

2.1.2 Support Vector Machine (SVM)

Support vector machines (SVMs) are a set of related supervised learning methods used for classification. This method was originally introduced by Vapnik in 1995 for two-class recognition problem (Vapnik, 1995).

Its basis is the Structural Minimization Principle. Defined in a vector space, it tries to find a decision surface that separates the data points of two classes. The decision surface in a linearly separable space is a hyperplane. Given training examples labeled

either "yes" or "no", a maximum-margin hyperplane is identified which splits the "yes" from the "no" training examples, such that the distance between the hyperplane and the closest examples is maximized. This idea can also be generalized to high dimensional space and data points that are not linearly separable.

The decision surface can be written as $\vec{W} \cdot \vec{X} - b = 0$. $\vec{X}$ is an arbitrary data point to be classified, and the vector for $\vec{W}$ and the constant b are learned from a training set of linearly separable data. Let D={ $(y_i, \vec{x_i})$} denote the training set where $y_i \in \{\pm 1\}$ is the correct classification result for $\vec{x}$ (+1 for being a positive example and -1 for being a negative example of the given class). The SVM problem is to find $\vec{w}$ so as to satisfy the following constraints:

$$\vec{w} \cdot \vec{x_i} \text{ -b >=+1 for } y_i = +1 \qquad (1)$$

$$\vec{w} \cdot \vec{x_i} \text{ -b <= -1 for } y_i = -1 \qquad (2)$$

The following figure illustrates the hyperplane that separates positive and negative examples.

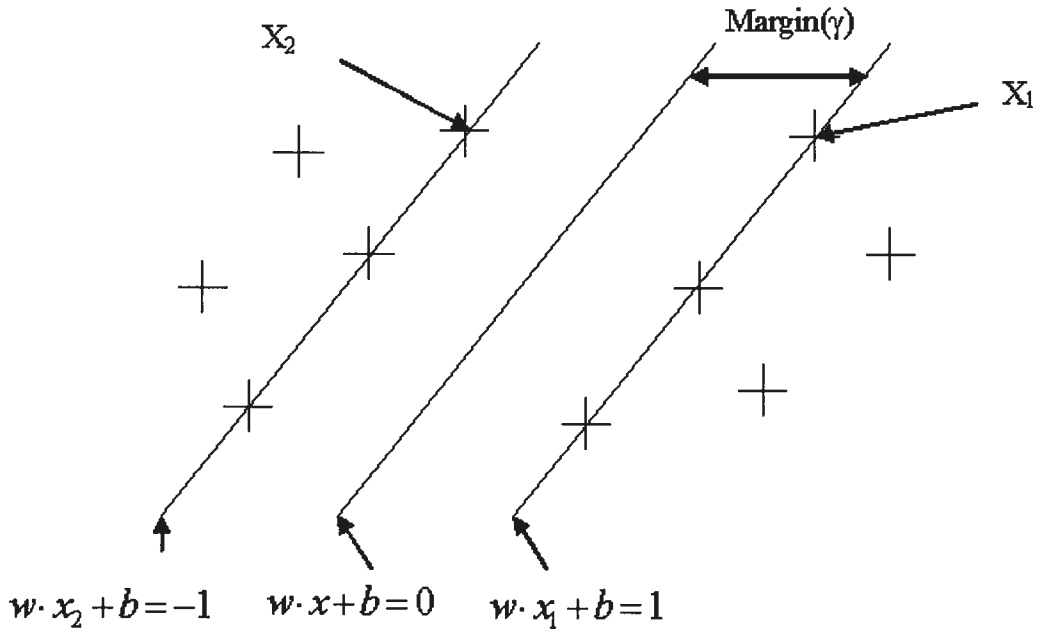

Figure 2.1: SVM Classification (McCulloch, 2004)

The parameters of the maximum-margin hyperplane are derived by solving a quatratic programming (QP) problem. There exist several specialized algorithms such as Platt's SMO algorithm that quickly solves the QP problem that arises from SVMs. The algorithms can also be extended for solving linearly non-separable cases by either introducing soft-margin hyperplanes or using the kernel-trick.

Hyperplane, illustrated by figure 2.1, is determined by the distance between the data points that have exactly the distance *1/w* from the decision plane. These data points are called support vectors and are the only elements that have an impact in the training set. Even if all other data points are removed, the algorithm will still learn the same decision function. This property sets SVM apart from other algorithms: with other algorithms, all the data points are needed to optimize the decision function.

When applied to text classification, previous studies show that SVM often classifier outperforms other classifiers. SVM generalizes well in high-dimensional feature space. It acknowledges text properties such as most relevant features and sparse document vectors. Also, most text classification tasks are linearly separable. All these properties make SVM a suitable classifier for document classification task.

2.3 Methods to improve the classification

The performance of the classification is crucial for users to access information correctly. The information needs to be classified quickly and accurately. There exist several methods to improve the result of classification. In text classification, two important methods are feature selection and term weighting.

2.3.1 Feature selection

Feature selection is a process commonly used in machine learning. It consists of selecting a subset of the features available from the data for application of a learning algorithm.

High feature dimensionality is one of the main obstacles in text categorization. There are 10000 features and more for even a medium-sized collection, which exceed the number of available training samples. The number of features is too high for many classification algorithms to handle. Therefore, it is necessary to reduce the number of features without sacrificing the classification accuracy. In a word, feature selection can help increase the classification efficiency.

Feature selection is also very helpful to increase the classification quality. Words extracted from a document are not always relevant to the main theme of the document.

There might be "noise" in the documents, which is the presence of non-relevant and repetitive data. This type of information can often degrade the performance of many classifiers. Previous studies have shown that feature selection can improve the classification performance of many algorithms including Naïve Bayes, KNN, neural networks (Yang et al, 1997). In our application, calls for tenders usually contain much procedural information, which is the information on the submission procedure of the tender, and is non-relevant to the subject of the call for tenders and can be seen as noise in the document.

Many feature selection techniques have been developed and they have produced good classification results. Most of the automatic feature selection methods include the removal of non-informative terms according to corpus statistics and construction of new features which combine low level features into high level orthogonal dimensions (such as the features extracted from Latent Semantic Analysis – LSI). The most popular one is the technique of information gain.

2.3.1.1 Information Gain

Information Gain (InfoGain) is a measure based on entropy. Entropy is a measure of the expected amount of information conveyed by an as-yet-unseen message from a known set (Quinlan, 1993).

The amount of information conveyed by a message, in bits, is the negative base-two logarithm of the probability of that message. For example, if there are 8 equally probably messages, receiving any one of them conveys $-log_2(1/8) = 3$ bits of information. Less probable messages convey more information, and vice versa. The expected amount

of information conveyed by any message is simply the sum over all possible messages, weighted by their probabilities.

In the context of supervised learning, the possible "messages" are prediction of the classes into which the data fall, and the probability of a class is the portion of the training cases which is labeled with that class. The entropy of the training set is the expected amount of information conveyed by the label of a case. A training set evenly split across the classes therefore has maximal entropy (1 if there are two classes), while one containing only examples of one class has entropy 0.

If the data can be divided into subsets by some useful test (i.e., by examining one of the features), each subset will have less entropy than the whole set. The *split* ***entropy*** for some feature is the sum of the entropies of the subsets resulting from the split, weighted by their sizes as fractions of the size of the original set.

The information gained by splitting on some feature is simply the original entropy minus the split entropy of the feature. In growing a decision tree, the feature offering the greatest information gain is selected at each step.

More precisely, given a set E of classified examples and a partition $P\{ E_1, ..., E_n \}$ of E, the information gain is defined as

$$G(E, P) = entropy(E) - \sum_{i=1,...,n} entropy \quad (E_i)^* \mid E_i \mid / \mid E \mid$$

Intuitively spoken the information gain measures the decrease of the weighted average impurity of the partitions $E_1, ..., E_n$, compared with the impurity of the complete set of examples E.

In document classification, information gain measures the number of bits information obtained for category prediction by knowing the presence or absence of a term in the document.

If $\{C_{i\}}\}_{i=1}^{m}$ is a set of categories, the information gain of a term t can be defined as

$$G(t) = -\sum_{i=1}^{m} P_r(c_i)\log P_r(c_i) + P_r(t)\sum_{i=1}^{m} P_r(c_i \mid t)\log P_r(c_i \mid t)$$

$$+ P_r(\bar{t})\sum_{i=1}^{m} P_r(c_i \mid \bar{t})\log P_r(c_i \mid \bar{t})$$

If we have a corpus, we can compute the information gain for each unique term. Then we can remove those features whose information gain is less than a predefined threshold, or keep the k terms with the highest information gains. The computation includes the estimation of the conditional probabilities of a category given a term, and the entropy computations in the definition. The probability estimation has a time complexity of $O(n)$ and a space complexity of $O(VN)$ where N is the number of training documents, and V is the vocabulary size.

The information gain method is one of the widely used feature selection techniques. Also previous studies have shown that among different dimensionality reduction (a.k.a. feature selection) techniques, information gain is one of the most effective (Yang et al, 1997).

2.3.1.2 Other feature selection methods:

Other statistical feature selection methods include Document Frequency (DF) thresholding method, mutual information (MI) method, χ^2 statistic (CHI) method, Term Strength (TS) method (Yang et al, 1997).

There are other non-statistical feature selection techniques available. Sentence filtering is one selection method that takes place prior to the indexing process of the training phase. It consists of rejecting or retaining sentences in a document based on the importance of the sentences in the document. There has been some early work on passage filtering based on character n-grams (Cavnar, 1993). The automatic summarization (Orasan et al, 2004) is one technique used to identify the most meaningful sentences. For example, the relevancy of a sentence can be estimated based on its position, length, the frequency of the terms and its similarity with the title. Named entities are also used in some of the studies related to text classification. However, in most cases, the use of named entities in classification is restricted to replacing common strings such as dates or money amounts with tokens corresponding to the class name, to increase the ability of the classifier to generalize. In the project in which I participate – MBOI, one previous study used named entities as indicators of relevance of passage in the filtering process (Paradis, 2005). However, the classification result didn't show significant increase.

For conventional method like Naïve Bayes, feature selection usually improves classification accuracy of classifier and avoids "overfitting" (Yang et al, 1997). Compared to Naïve Bayes, SVM has the ability to generate well in high dimensional feature spaces (Joachims, 1998), since SVMs use overfitting protection that does not

depend on the number of features. Generally the performance improvement of SVM classifier using feature selection method hasn't been proved to be significant in terms of classification quality. However, it can still reduce the complexity of the calculation.

2.3.2 Term Weighting Methods

Term weighting consists of adjusting the term weight of the vector space model in the document indexing process. In vector space model, a document is represented as a vector of features using Term Frequency (TF) and Inverted Document Frequency (IDF). This model simply counts TF without considering the importance of the sentence in which it appears. As we have shown in our earlier example, sentences in a document have different importance for identifying the content of the document. Thus, by assigning a different weight according to the importance of the sentence, we can achieve better results.

2.3.2.1 Document Indexing and Term weighting

In the indexing phase, the features of the documents are extracted and indexed. The indexed document is commonly represented using the vector space model (Salton et al., 1975). This model represents documents as a vector whose components are all the possible index terms. Each index term has an associated weight that indicates the importance of the index term in the document (or query). The weight is usually represented by combining two factors: the importance of each index term in the document and the importance of the index term in the collection. The importance of each index term can be measured by the number of times that the term appears in the document. This is called the *term frequency* which is denoted by the symbol *tf*. The

importance of the term in the whole set of document is represented by the rarity of the index term in the whole set of document. For example, an index term that appears in every document in the collection is not very useful, but a term that occurs in only a few documents may indicate that the term best represents the document in which it appears. This factor is usually called the *inverse document frequency* or the *idf factor.*

Mathematically the inverse document frequency (idf) can be expressed as:

$$idf = \log \ N/N_i$$

where N= Number of Documents in the collection;

N_i = Number of documents that contain the term.

By combining these two factors, we can obtain the weight of an index term *i* as:

$$W_{i,j} = tf \times idf$$

$$= tf \times log \ N/N_i$$

For terms that are relevant, we would like to assign a high weight $W_{i,j}$ and for terms that are considered irrelevant, we assign a lower weight. However, the above *tf*idf* is not always enough. So, a few alternative methods have been proposed.

2.3.2.2 Other term weighting methods

In previous studies surveyed, most of the term weighting methods measure the relevance of the terms regardless of the location of the terms in the documents. A term appearing in the title of a document is considered in the same way as a term in the body of the document. However, some studies consider the position information of the terms.

The title is often indicator of relevant terms. (Mock et al., 1996) is a study which used the title of a document in order to choose the relevant terms. The terms in the subject of the news articles were assigned highest weight, followed by terms in the text body section. The terms in the author section are assigned the lowest weight. The result shows an improvement of classification result.

Another study used two kinds of text summarization techniques to assign weight to terms. One method assigns higher weights to the sentences that are more similar to the title. Another method first measures the importance of terms by TF, IDF, and $\chi 2$ statistic values. Then we assign a higher importance to the sentence with more important terms. Both methods proved to make significant improvement on Naïve Bayes, SVM, Rocchio, KNN classifier (Ko et al, 2002).

2.3.2.3 Summary of Classification Approaches and Description of Our Application

We concur that automatic classification is a process that organizes information and helps to make information useful to everyone. The performance of the classifier is crucial because it allows users to get information quickly and accurately. The performance of the classifiers can be improved by the feature selection and term weighting methods. Both methods rely on techniques of distinguishing relevant and irrelevant features. For feature selection, most existing techniques use corpus statistics to distinguish between informative terms and non-informative terms. For term weighting, most of the methods measure the relevance of the terms using word frequency regardless of other factors.

Even though there have been a few alternative term weighting methods that exploit the position information, there have not been extensive investigation on alternative factors other than term frequency. In this thesis, we will investigate several alternative factors.

Our application problem is the classification of call-for-tender documents (CFT) downloaded from the web, according to one of the existing norms such as NAICS (North American Industry Classification System). The ability to classify call for tender documents into different domains of activity allows suppliers (users) to find the ones that are relevant to their business. As we showed earlier, an important problem in CFT classification is that this type of documents often contains procedural information irrelevant to the main theme. It's very difficult to eliminate this type of features effectively by using standard feature selection and term weighting methods: the irrelevant terms can appear quite frequently in a document, so the standard methods would assign quite high weights to these terms.

We also showed in the earlier example that the position of terms in the document, certain concepts and named entities extracted from calls for tenders can be indicators of the irrelevant procedural part or the relevant part of a CFT. In this thesis, we will propose feature selection methods and term weighting methods using additional information such as position information, named entities and concepts to improve the performance of classifiers. In the next chapter, we will describe the factors we use in this study.

CHAPTER 3

RELEVANCE INDICATORS

Calls for tenders often contain lengthy procedural information unrelated to the subject in the document.

The following is another example of call for tender document:

"CITY OF GREATER SUDBURY EXPRESSION OF INTEREST FOR COMPUTER HARDWARE FOR POLICE SERVICES The Greater Sudbury Police Services Board requires a vendor of record to supply a quantity of rugged mobile computers for use in Police Vehicles. *Expression of Interest packages may be obtained at the City of Greater Sudbury, Supplies Services Department, Main Floor, Tom Davies Square, 200 Brady Street, Sudbury, ON P3A 5P3 OR may be downloaded from our website at www.city.greatersudbury.on.ca. If you download the document from our website, please ensure you sign up to receive addendums. If you do not sign up, you will not receive addendums. Your expression of interest package must be returned to the City of Greater Sudbury, Supplies Services Department, Main Floor, Tom Davies Square, 200 Brady Street, Sudbury, ON, P3A 5P3 NO LATER THAN 11:00 a.m. (our time), Friday, June 6th, 2003, using one of the methods indicated in the expression of interest document. Questions regarding any aspect of this expression of interest must be forwarded by e-*

mail to stephanie.cundari@city.greatersudbury.on.ca or by fax to (705) 671-0871. The deadline for questions is 4:30 p.m. on Tuesday, June 3rd, 2003. Addendums must be issued 48 hours prior to the close of this expression of interest. Answers to questions will be sent to all bidders in an addendum format. The City reserves the right to accept or reject any or all Expressions of Interest."

The part of document in italics is the procedural information that is unrelated to the subject of the document. From the document above, we see that the subject is often introduced in the first few sentences of the document. Therefore most of the noise is present in the later part of the document. So the first intuition is to make use of the position information to detect if a sentence contains useful information for the classification purposes.

Also, we see that some named entities are also very important in distinguishing relevant and irrelevant information. Named entities such as email and URL present the point of contact of contracting authority which is part of the procedure submission information unrelated to the subject. E.g. "Questions regarding any aspect of this expression of interest must be forwarded by e-mail to stephanie.cundari@city.greatersudbury.on.ca". In the same way, there are a number of other types of named entity appearing in the irrelevant parts of CFT, such as DATE, LOCATION, and so on. If a sentence contains such named entities, it has a high chance to be irrelevant for classification.

Another important factor is concept. A concept is an abstract or generic idea generalized from particular instances(Merriam-Webster "concept", 2006). Specialized

concept often presents product services, which are subjects of the document. For example, the sentence "Greater Sudbury Police Services Board requires a vendor of record to supply a quantity of rugged mobile computers for use in Police Vehicles." contains the specialized concept "rugged mobile computers", which is related to the subject of the whole document. We observe that sentences containing such concepts are usually relevant sentences.

The sample call for tender document shows that position information of the features, presence of concepts and named entities are three factors that can help to distinguish relevant features and irrelevant features in a document. All three factors can be used in sentence filtering and term weighting methods to improve the performance of the classifiers. In sentence filtering, we can use those factors to determine whether a sentence is important or not. In term weighting, we can use those factors to determine the how important the sentence is and then modify the weighting of terms in the sentence accordingly.

Before describing in detail how we use these additional characteristics in our method for term selection and weighting, let us first describe the common methods used to extract the characteristics we require.

3.1 Named Entities

3.1.1 Information Extraction overview

Information Extraction (IE) is a language-engineering process that extracts structured or semi-structured data determined by a set of pre-defined extraction criteria. The data can then be used to be displayed to users or used for other application. As an

application for information extraction, for example, IE system can scan a set of unstructured written resumes, automatically extract the name, address field of each applicant. Those fields can then be inserted into database.

Currently, there are five types of information extraction: Named Entity recognition (NE), Coreference Resolution (CO), Template Element construction (TE), Template Relation construction (TR) , and Scenario Template production (ST) (Grishman & Sundheim, 1996). The task of NE recognition is to find entities in the document such as organizations, locations, persons, date/time, money/percentage, etc. The task of CO is to identify relations between entities in texts. For example, for the sentence "poor David, I pity him!!" coreference resolution would tie "David" with "him". TE is used to add descriptive information to NE results using CO. For example, after NE identifies Shanghai as a location, TE will add the information that this is in China. TR is to find relations between TE entities. ST is to Fit TE and TR results into specified event scenarios. For example, TE will identify Jacques Cartier as a French explorer and Quebec as a city of Canada. ST will identity facts like Jacques Cartier first discovered Quebec in 1534. In this thesis, only Named entities recognition is used for the research because NE is the indicator of the relevance of a sentence to the subject of a document. Other information extraction processes are not used.

Information extraction system is evaluated in Message Understanding Conference (MUC). Message Understanding Conference is a DARPA sponsored conference in which participating IE systems are rigorously evaluated. Information extracted by the systems from blind test sets of text documents are compared and scored against information manually extracted by human analysts (Grishman & Sundheim, 1996).

3.1.2 Named Entity Recognition

Named entity recognition (NER) is an information extraction task that extracts from unstructured documents information of predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. For example, a NER system producing MUC-style output might tag the sentence as following,

Jim bought 300 shares of Acme Corp. in 2006.

<ENAMEX TYPE="PERSON">*Jim****</ENAMEX>*** *bought* ***<NUMEX TYPE="QUANTITY">****300****</NUMEX>*** *shares of* ***<ENAMEX TYPE="ORGANIZATION">****Acme Corp.****</ENAMEX>*** *in* ***<TIMEX TYPE="DATE">****2006****</TIMEX>****.*

There are two major approaches to NER. One uses linguistic grammar rules and the other one is based on statistical model. The grammar rule-based systems typically obtain better results, but takes lots of manual work from linguists. In contrast, the statistical NER systems require much training data, but can be ported to other languages much more rapidly and require less work overall(Wikipedia "NER", 2006).

In the following section, I will describe the NER system GATE, the system used in our NE experiments.

3.1.3 GATE information extraction system

The NER system used in this thesis is GATE ANNIE system. GATE system is an infrastructure developed by researchers at the University of Sheffield since 1995 for developing and deploying software components that process human language. It has been used in a wide variety of research and development projects. ANNIE(A Nearly-New IE) is the information extraction component of GATE. It is developed by Hamish Cunningham, Valentin Tablan, Diana Maynard, Kalina Bontcheva, Marin Dimitrov and others from the natural language group of the University of Sheffield. It is an open-source, robust Information Extraction (IE) system based on finite state automates. It consists of the following main language processing tools: tokenizer, sentence splitter, Gazetteer lists, POS tagger, semantic tagger (Cunningham et al, 2006). All the components form a pipeline that take document corpus as input and output extracted named entities.

First, the tokeniser splits text into simple tokens, such as numbers, punctuation, symbols, and words of different types (e.g. with an initial capital, all upper case, etc.). Then the gazetteer lists are searched to find all occurrences of matching words in the text. The gazetteer lists used are plain text files, with one entry per line. Each list represents a set of names, such as names of cities, organizations, days of the week, etc.

Below is a small portion of the list for units of currency:

```
Ecu
European Currency Units
FFr
Fr
```

```
German mark
German marks
New Taiwan dollar
New Taiwan dollars
NT dollar
NT dollars
```

The sentence splitter splits the text into sentences. This module is required for the POS (Part-of-speech) tagging. The POS tagger is a modified version of the Brill tagger (Brill, 1993), which produces a part-of-speech tag as an annotation on each word or symbol. Neither the splitter nor the tagger is a mandatory part of the IE system, but the extra linguistic information they produce increases the power and accuracy of the IE tools.

The semantic tagger consists of pattern-action rules, executed by the finite-state transduction mechanism. It recognizes entities like personal names, organizations, locations, money amounts, dates, percentages, and some types of addresses.

The ANNIE system supports multiple languages through Unicode.

ANNIE can be used and customized in GATE's graphical development environment and integrated in other applications through its API (Application Programming Interface).

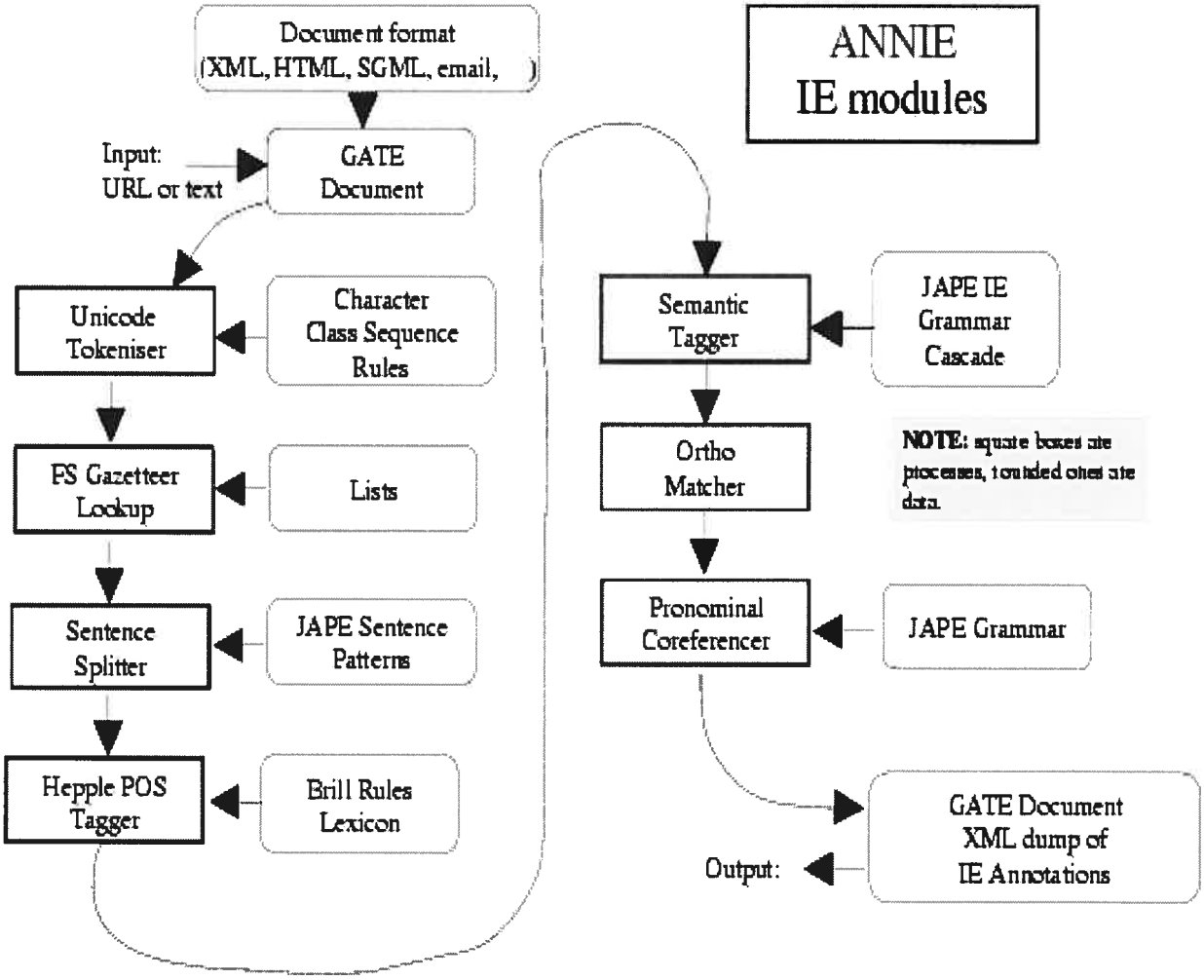

Figure 3.1: Annie NE extraction Process (Cunningham et al, 2006).

The GATE ANNIE named entity recognition system has been used in the MUC-6 Named entities recognition task and many other experiments on question answering and text summarization. In MUC-6, Location, Organization, Money, Date, Time, Person and Percent are extracted for the task. The official score shows that the precision of the system is 94%, the recall is 84% and the F-measure is 89%. (Gaizauskas et al, 1995). Given that human annotators do not perform to the 100% level (measured in MUC by

inter-annotator comparisons), NE recognition can now be said to function at human performance levels.

3.1.4 Named Entities in Call for tender documents

CFT (Call for tender) documents are semi-structured documents. They don't have a standard format. However, each CFT document always contains standard information such as contracting authority, delivery and closing date of the CFT, URL and email address of the contract authority, etc. This type of information is examples of named entities that can be extracted using information extraction system.

By analyzing the type of information named entities in CFT documents contain, we can predict whether or not NE or the part of the document containing NE is related to the subject of the document. In this thesis, we consider the following named entities:

1. Organization:

   Organization is usually the contracting authority of the CFT document. The organization, when located at the beginning of the CFT document, can be an indicator of the subject of the CFT document. eg. CITY OF GREATER SUDBURY EXPRESSION OF INTEREST FOR COMPUTER HARDWARE.

2. Location:

   The location is the location of the delivery place or the location of the contracting authority.

3. Date:

It can be the opening or the closing date of the call for tender. Also the delivery date is a possibility.

4. Money:

The value indicates the contract or the business size required in the call for tenders.

5. Person:

The contact person for the call for tender.

6. URL:

This is the URL of the call for tender document or the URL of the contracting authority. It can be strong indicator of the non-relevance of the part of the document containing it, especially if it is located towards the end of the document. In this case, it often refers to the standard point of contact of the CFT without revealing the subject of the document.

7. Email:

This is the email address of the contracting authority. It can be a strong indicator of non-relevance for the same reasons as for URL.

3.2 Concept Extraction

Similar to the NE, another factor whose presence can indicate the relevance of the sentence is concept.

3.2.1 Concept Definition

A concept is an abstract or generic idea generalized from particular instances(Merriam-Webster "concept", 2006). Concepts can be categorized into simple concepts and complex concepts. Simple concepts are concepts that consist of only one word (eg. statement, software, president, etc). Complex concepts have more than one word (eg. financial statement, management software, US president, etc).

Since the concepts present ideas of the document, they can be seen as positive indicators of relevance to the subject. Part of the document that doesn't contain concept can be considered to be less relevant.

3.2.2 Concept Extraction Techniques

There are many concept extraction systems. Most systems are rule-based systems including the *Nstein NConcept Extractor* that is used in my research. Because it is a commercial product, the detail of its concept extraction method is not available. But we know it is an extraction system that uses a combination of linguistic-based and statistics-backed processes to locate and retrieve concepts (Lemay, 2006). In the academic field, some simple systems extract concepts by detecting words with high frequency in the document - i.e. high *tf* (term frequency). Alternatively, other methods extract concepts by detecting words with high value of *tf-idf* (term frequency and inverse document frequency), which corresponds to detecting words that not only have high frequency in the current document but also are relatively rare in a large collection of documents. More complex concept extraction methods can involve for example the method of Lin. Lin used a concept generalization taxonomy in the form of WordNet. For example, If "laptop" or "hand-held computer" are found in a text, we can infer that the text is about

"portable computer", which is the parental concept. Then if "mainframe computer" is mentioned in the text, the concept "digital computer" , which is the parental concept of both "portable computer" and "mainframe compter", can also be inferred. In order to find the most appropriate concept generalization, Lin used a ratio-based method to determine cutoff points to retain the appropriate level of generality for the important concepts extracted by WordNet. The details of Lin's method can be referred to (Lin, 1995).

The tool to extract concept in the thesis is *Nstein NConcept Extractor*, a tool based on text mining linguistic and statistic processes (Lemay, 2006). The following example shows the output of a NER by this tool:

> "Landfill Disposal Services. Landfill Disposal Services for <**Complex**>solid waste</**Complex**> generated and delivered by various Navy installations within Navy Region Northwest. Contractor shall accept and dispose of <**Complex**>solid waste</**Complex**> delivered to facility. Commodities types include but not be limited to, municipal <**Complex**> solid waste<**Complex**>, creosote pilings, <**Simple**>soil<**Simple**>, <**Simple**>concrete, <**Simple**>asphalt, construction demolition and land clearing debris, sludge cake/filter press dewatered solids, petroleum contaminated soils and truck wastes. Landfill <**Complex**>service provider</**Complex**>shall ensure compliance with all federal, state, and local laws or regulations related to refuse disposal and can be licensed by Washington State Utility and Transportation Commission. Electronic monthly billing services to include record of Daily transaction detail by

date, time, net weight in tons, commodity delivered and vehicle <**Complex**>identification number</ **Complex**>."

The tags <Simple> and <Complex> respectively mean Simple and Complex Concept. This is the system we will use in this study.

3.3 Position Information

Similar to the presence of concept information, position information of the features can also indicate whether the feature is relevant to the subject of the document.

In the research of automatic summarization, documents are often split into sentences and sentence position is considered to be very important to determine whether the sentence can be part of the summary of the document. One important sentence heuristic assumes that relevant sentences are generally sentences lying at the beginning and the end of a document, in the first and last sentences of paragraphs, and also immediately below section headings. This heuristic using position gave very good results in several summarization experiments (Edmundson, 1969; Kupiec, 1995).

Alternatively to sentence selection, term weighting can also benefit from the position information of the sentence: We can assign weights according to the position of the sentence in which words appear. For example in the study by Mock (Mock et al., 1996), the terms in the subject of the news articles were considered to be the most relevant and assigned the highest weight, followed by terms in the text body section. The terms in the author section are assigned lowest weight. The result shows an improvement of classification result.

In the call for tender document, as shown as an earlier example in this chapter, the subject of the tender is often introduced in the first few sentences. The last few sentences usually give information on procedures of tender submission. Therefore, the first few sentence of a paragraph will be considered to be more relevant than the last few sentences.

In the next chapter, we will describe the experiment that tries to reveal the utility of the three factors as indicator of relevance and then discuss the result. We will also describe the approaches to exploit the three types of information we introduced. Two alternative methods will be proposed: using these types of information to select important sentences and filter out the other sentences, or using them to modify the traditional term weighting methods.

CHAPTER 4

USING SPECIFIC FEATURES OF CFT TO IMPROVE TERM WEIGHTING

In chapter 3, named entities, concepts and position information are presented as three specific features that are susceptible to determine the relevant part of the document.

In section 4.1 of this chapter, I will describe the experiments conducted that further confirm the utility of each of the three factors to determine relevant sentences. Then in section 4.2, I will describe the feature selection and term weighting methods that incorporate these factors to improve the performance of classifiers.

4.1 Identification of Important Sentences by Specific Features

4.1.1 Identification of Important Sentences by position

It's very important to understand how useful position information is and determine the position of the most relevant sentences of the call for tender documents.

Those are the purposes of this experiment. Then we can conduct classification experiments by either filtering out irrelevant sentences or increasing the term weights of features in relevant sentences.

A set of 1000 sentences from different documents has been selected and manually annotated as being important (Y) or non-important (N). This set of 1000 sentences were selected by Francois Paradis and featured in his classification experiment(Paradis, 2005).

The position of each sentence is studied to evaluate the part of the document that with the most relevant sentences. Table 1 presents the accuracy of the relevant sentences for each section of the document, which indicates percentage of relevant sentences in this section of the document. For example, if there are 334 sentences in total in this section of the document and there are 131 relevant sentences. The accuracy of relevant sentences is 131/334, 39.34%. Also the percentage of all the relevant sentences, which measures the percentage of relevant sentences of this section over all the relevant sentences, is presented in the table.

	% of all the relevant sentences	**accuracy of relevant sentences**
First 1/3 (334)	53.91% (131/243)	39.34% (131/334)
1/3 to 2/3(333)	30.45% (74/243)	22.22% (74/333)
Final 1/3 (333)	15.64% (38//243)	11.41% (38/333)

Table 4.1: Identification of important sentence according to its position

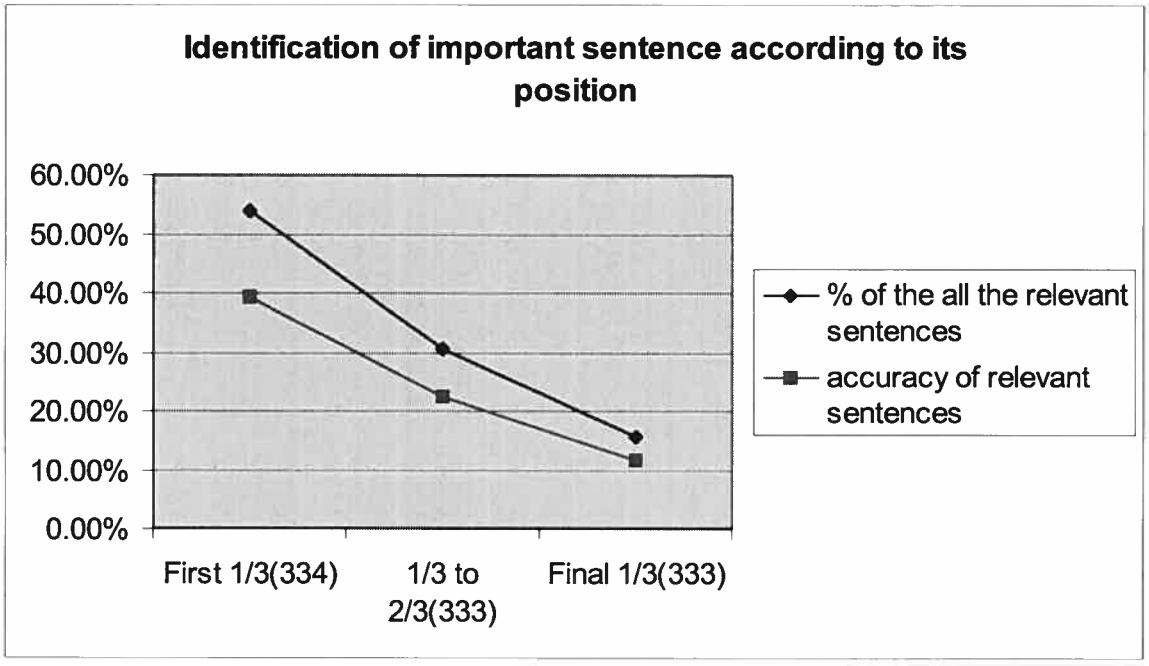

Figure 4.1 Identification of important sentence according to its position

The result shows the percentage of relevant sentences and the accuracy of relevant sentences for first 1/3 sentences, from 1/3 to 2/3 and the final 1/3 sentences of each the call for tender document. We see that the first 1/3 of the document contains most relevant sentences (53.91%) and the final 1/3 has the least number of relevant sentences (15.64%). Also, out of the 334 sentences of the 1/3 of the document, 39.34% are relevant compared to only 11.41% for the last 1/3 of the document. Therefore, this preliminary test shows that the first 1/3 of sentences of the document contain more relevant sentences than the rest of the document.

4.1.2 Identification of Important Sentences by Concepts

The purpose of this experiment is to see whether concepts relevant to subject help to identify relevant sentences. First, in the section 4.1.2.1, I will explain the process of extraction of relevant concepts in call for tenders. Then in the section 4.1.2.2, I will use an experiment conducted to show how relevant CFT sentences containing extracted concepts are.

4.1.2.1 Relevant Concept Extraction

Concepts present ideas of the document. Therefore, if a sentence contains a concept, it's susceptible to be more relevant than the other sentences. However, not all concepts extracted are relevant to the subject. Only concepts related to subject of the documents are relevant for the classification process. In this thesis, the North American Industry Classification System (NAICS) presents the class to which calls for tender are classified to. NAICS is defined as a classification code system that lists sectors of economic activity organized in a hierarchical style.

Here is an example of several codes inside NAICS definition:

562 Waste Management and Remediation Services

5622 Waste Treatment and Disposal

56221 Waste Treatment and Disposal

562211 Hazardous Waste Treatment and Disposal

562212 Solid Waste Landfill

Also for each code, NAICS definition document contains its description. For example, for code 562211 "Hazardous Waste Treatment and Disposal ", its description is

"This U.S. industry comprises establishments primarily engaged in (1) operating treatment and/or disposal facilities for hazardous waste or (2) the combined activity of collecting and/or hauling of hazardous waste materials within a local area and operating treatment or disposal facilities for hazardous waste... sewer systems or sewage treatment facilities)--are classified in U.S Other Nonhazardous Waste Treatment and Disposal."

Therefore, if a concept matches closely to the sectors of economic activities listed in NAICS definition, it is very possible that it is relevant to the subject of the call for tender (CFT) document. In order to see how closely matched a concept and the description of each NAICS sector is, we can use the IR similarity score that is generated after doing a information retrieval on each NAICS description using concept. High score means the concept is very related to this particular NAICS product and service category and low score means it is not. A threshold value is used to distinguish the high and low score.

To further explain how concept relevant to subject are extracted, the following concept extraction steps are presented below:

1. For each CFT document, extract concepts (Simple Concepts, Complex Concepts) using Nstein's NConcept Extractor, which is a tool for concept extraction described in more detail in chapter 3.

2. Filter concepts with:

2.1 Stop list removal:

Stop List contains a list of concepts are concept that are often present in many documents but irrelevant to the subject of the document. The list contains concepts such as "statement", "contract", "procurement", "competition", "provider", "provision", etc. They are examples of the stop list. Removal of the stop list words cleans up the noise.

3. Match with NAICS definitions using Information Retrieval

It is our intuition that the more closely the concepts extracted is matched to the NAICS definition, the more relevant the concept is to the subject. In order to achieve this, we can do an information retrieval (IR) of the NAICS Definition documents using the concepts extracted. This IR approach is used because we can easily evaluate the relevance of the concept to the subject by using the IR similarity score between the concept and each NAICS code description. Therefore for the IR step, the query of the IR is the concept extracted after Step 2 and the document searched is NAICS code definition.

For each concept, we obtain an IR similarity score for each document of the NAICS definition. Each document of NAICS definition describes a sector of economic activity. IR score ranges between 0 and 1.0. The higher the IR score means the concept is more closely matched to one sector of economic activity and therefore can be seen as relevant.

The feature selection method using concept information consists of selecting the sentences that contain concepts extracted using the process described above and filtering out sentences that don't contain those concepts. For example, we can keep all the concepts that contain a document IR score over 0.8.

As an example, here is a CFT document:

Landfill Disposal Services . Landfill Disposal Services for solid waste generated and delivered by various Navy installations within Navy Region Northwest. Contractor shall accept and dispose of solid waste delivered to facility. .....Landfill service provider shall ensure compliance with all federal, state, and local laws or regulations related to refuse disposal and can be licensed by Washington State Utility and Transportation Commission. Electronic monthly billing services to include record of Daily transaction detail by date, time, net weight in tons, commodity delivered and vehicle identification number.....

Two concepts extracted by NConcept Extractor are "solid waste" and "monthly billing". After searching on NAICS definition, the IR score of "solid waste" on the document below is 1.0.

562111 Solid Waste Collection

This U.S. industry comprises establishments primarily engaged in one or more of the following: (1) collecting and/or hauling nonhazardous solid waste (i.e., garbage) within a local area; (2) operating nonhazardous solid waste transfer stations; and (3) collecting and/or hauling mixed ...

But the IR score for "monthly billing" on NAICS definition returns 0. Therefore, we only keep "solid waste" as the final concept. We can successfully determine the useful concepts in the sentences in this way.

4.1.2.2 Concepts as Indicators of Important Sentences Experiments

In this experiment, the same 1000 sentences for experiments described in section 4.1.1 are used. They are first tagged as either relevant or relevant manually. Then the 1000 sentences go through the extraction of concepts described in 4.1.2.1. At the end of the concept extraction process, the concepts that have IR score over 0.8 are kept and otherwise rejected. The sentences that contain at least one concept-NAICS are determined to be relevant. We then try to compare the relevant sentences determined manually and the ones assigned using concepts to see if sentences containing concepts are positive indicators and sentences not containing concepts are negative indicators.

Table 4.2 shows the accuracy of the extracted concepts as positive indicators on the 1000 sentences sub-collection of FBO. We see that the absence of concepts is a very strong indicator of irrelevant sentences (88%). Sentences with concepts don't give an accurate result: 166 sentences are relevant among all 279 sentences extracted through the Nconcept Extractor, which gives an accuracy of 59%. We found that certain concepts, such as "contract" and "service" are not really relevant to the subject of each document.

	Type	Accuracy
Sentences with Concepts	positive indicator	166/279 (59%)
Sentences without Concepts	negative indicator	637/721 (88%)

Table 4.2: Concept Type and Accuracy

4.1.3 Identification of Important Sentences by NE

Each type of named entities can be seen either as negative indicators (N) or positive indicators (P) of relevant sentence to the subject. The purpose of this experiment is to see which named entities are negative/positive indicators and how accurate each named entity can indicate the relevance of each sentence. Experiments are done on a corpus with 1000 sentences. The named entities and the relevant sentences are manually tagged.

The table below shows the accuracy of the entities as positive/negative indicators on the 1000 sentences subcollection of FBO, the same collection used in experiments in section 4.1.1. For example, URL (a negative indicator) appeared in 15 sentences, 15 of which were labeled negative. The accuracy is 100 %.

From the table, we see the URL, Money and Email are very strong indicators of irrelevance of the sentences. Organization is a positive indicator of relevant sentence. Other named entities give mixed signals on the relevance of the sentences.

NE Types	Relevance Type	Accuracy
Organization	P	(125/190) 66%
Location	N	(47/94) 50%
Person	N	(80/122) 66%
Date	N	(106/140) 76%
Money	N	(17/17)100%
URL	N	(15/15) 100%
Email	N	(20/20) 100%

Table 4.3: NE Experiments on identification of important sentences

Therefore, we choose to use organization, URL and Email in the term weighting and feature selection methods because they are strong indicators of relevance of the sentence.

As we see through the experiments in 4.1, NE, concepts and position information can indicate accurately the relevance of each sentence in the document. In section 4.2, we will introduce term weighting and feature selection methods and describe how we can use three factors in those methods.

4.2. Term Weighting and Feature Selection Methods

The text classification process normally consists of a training phase and a text classification phase. The training phase consists of the preprocessing and indexing of the documents. At the end of this step, a model is built containing the statistics of the class. In the classification step, the model built is used to perform the classification. The training phase can be illustrated in figure 4.1.

As we explained earlier in the thesis, two methods, feature selection and term weighting that take place in the training phase of the classification can be used to improve the classification result. Traditional feature selection and term weighting methods rely on statistical methods to determine relevant features. However, as explained in chapter 3, the traditional approaches can't always determine accurately the relevant and irrelevant features of the document. By incorporating the additional information such as sentence position information, presence of named entities and concepts related to subject of the document into the feature selection and term weighting method, we hope to achieve better classification results than the traditional methods.

4.2.1 Methods to improve classification results

We propose two types of approach: selecting a subset of features according to the relevance of sentences; weighting features according to the relevance of sentences.

4.2.1.1 Feature Selection Methods

Document preprocessing step is the first step of the training phase. In this step, the documents are segmented into sentences. The documents in our research are calls for tenders. A call for tenders usually consists of some meta-data such as the date of

publication (21 May 2001"), classification codes (NAICS \424120" and FCS \75"), the contracting authority ("Environmental Studies"), etc. The body of the document is composed of the subject line and the description; only these fields will be used for document preprocessing.

Feature selection takes place after the sentence splitting step. In our case, it consists of filtering out sentences that are considered to be irrelevant to the subject and keeping those sentences that are relevant. Sentence is the unit of passage to be evaluated because it is a basic unit of language that often represents an idea. It is also easy to segment the document into sentences.

As seen earlier in this chapter, named entities, position and concept information are factors that can possibly distinguish relevant features and irrelevant features. We can use those factors in feature selection method.

4.2.1.2 Term Weighting Method

As explained in chapter 2, during the training phase of the classification process, each document is indexed using vector space model. Each indexed term has an associated weight that indicates the importance of the index term in the document (or query). The weight is determined by the term frequency and the inverted document frequency of each term. In my term weighting method, the unit of passage to be evaluated for relevance is sentence. We also know also that there are relevant sentences and irrelevant sentences in the document to the subject. Therefore, we can assign more weight to terms in sentences

more relevant to the subject and less weight to terms in sentences less relevant to subject. By doing so, we hope to achieve better result in classification.

In the original vector space model (Salton et al., 1983), the weight of the indexed term t_i in a document d_j is calculated by using this formula:

$$W_{ij} = tf_{ij} \times idf_i$$

$$= tf_{ij} \times log\ N/N_i$$

where tf_{ij} = term frequency of t_i in document d_j

idf_i = Inverted Document Frequency

N = Number of Documents in the collection

N_i = Number of documents that contain the term

In our study, we use a simple method to incorporate the relevance of sentences: We modify the *tf* value of the term t_i by multiplying it by the relevance factor of the sentence in the document. That is, the new term weight is determined as follows, where *RF* is a function measuring the *relevance factor* of a sentence S_k in which the term occurs:

$$W_{,ij} = (modified\ tf_i) \times idf_i$$

$$modified\ tf_{ij} = \sum_{S_k \in d_j} freq(t_i) \times RF(S_k)$$

The modified *tf* value of a term t_i in a document d_j is the sum of the modified term frequency of t_i in each of the sentences of the document. We will describe in the next few sections how the relevance factor (RF) of each sentence is determined according to the

presence of named entity, presence of extracted concepts, sentence position or the combination of those three factors.

4.2.2 Feature selection and term weighting using named entities

Our proposed feature selection method keeps the sentences containing named entities that are positive indicators such as organization and eliminates sentences that are negative indicators, such as URL and email. In doing so, we can hopefully determine accurately relevant features and improve classification results.

For the proposed term weighting method, in order to achieve better classification results, we increase the term weights for features in sentences that contain positive indicators and decrease the term weights of features in sentences that contain negative indicators. For example, if organization is a positive indicator of relevance and URL is negative indicator of relevance, we can increase the term weights of sentences containing organization and decrease the term weights of sentences containing URL.

4.2.3 Feature selection and term weighting using concepts

The term weighting method using concept information consists of increasing the term weights of features of the sentences containing concepts. Since most relevant concepts can possibly have higher IR score by searching on NAICS code, we can increase more the term weights of the features in sentences containing concepts according to the IR score of the concepts. For example, the term weights of features in sentences containing concept with IR score over 0.8 is multiplied by a boosting factor. The term weights of features in sentences containing concept with IR score over 0.5 but less than 0.8 is multiplied by a boosting factor whose value is less than the boosting factor value

for the IR score over 0.8. The term weights of sentences with containing concept with IR score less than 0.5 can stay the same because by observation, most of those sentences are not related to the subject of the document.

4.2.4 Feature selection and term weighting using position information

Call for tender documents are most highly structured documents. The subjects of the documents are mainly presented in the early part of the document. The sentences in later part of the document often present irrelevant information such as the procedure of submitting a proposal or the contact information for the call for tender. Therefore, the position of each sentence is an indicator of the relevance of the sentence to the subject. Because of the sentences in early part of the document are more relevant to the subject of the document, the feature selection method consists of eliminating the sentences in the later part of the call for tender documents and keeping sentences in the early part of the document.

Another method to improve the classification result is the term weighting method. Previous research has shown by assigning term weight differently according to the location of the term in the document, the classification results can be improved (Ko et al, 2002). There are several ways to use position information in term weighting method. One approach is to increase the term frequency values of terms in sentences of the first few sentences of the document (eg. top 1/3 sentences of the document). In this thesis, I use an approach that determines a relevance factor that decreases along the way from beginning of the document to the end. This simple approach is used because as the sentence relevance to the subject decreases from beginning of the document to the end, the relevance factor value should also decrease, too. As an example, the term weights of

features in the first 1/3 sentences of the document can be multiplied by 8. The term weights of features in sentences between 1/3 and 2/3 sentences can be multiplied by 4. And the term weights of features in the last 1/3 sentences can stay the same.

4.2.5 term weighting method using combined factors

Concept information, named entities information and location information mentioned above are factors that can potentially influence the relevance of sentence in the document. We can assign a relative weight to sentences according to its relevance with concept information, named entities information and location information.

We can first select factors that are positive indicators or negative indicators of the sentence relevance. We can then increase the term weights of sentences containing positive indicators and decrease the term weights of sentences containing negative indicators to improve the classification results.

The method can be described by the following equations. First of all, our method consists of modifying the term frequency (tf) by a boosting factor $I(S_i)$. The modified term frequency can then be used in the classification indexing process.

$$tf' = tf * I(S_i)$$

The boosting factor $I(S_i)$ can be determined by the following equations.

$$I(S_i) = \begin{cases} 1/4 \Leftarrow url, email, money \in S_i \\ I_{pos}(S_i) + I_{Concept}(S_i) + I_{org}(S_i) \Leftarrow concept, org \in S_i, Pos(S_i) \le 1/3 * lengh(doc) \\ 1 \Leftarrow otherwise \end{cases}$$

$$I_{pos}(S_i) = \begin{cases} 1 \Leftarrow Pos(S_i) \le 1/3 * lengh(doc) \\ 0 \Leftarrow Otherwise \end{cases}$$

$$I_{concept}(S_i) = \begin{cases} 1 \Leftarrow Concept \in S_i \\ 0 \end{cases}$$

$$I_{org}(S_i) = \begin{cases} 1 \Leftarrow Organization \in S_i \\ 0 \end{cases}$$

Another way to describe how to determine the factor $I(S_i)$ is the following algorithm.

Algorithm:

$I(S_i)$ = 1

If sentence is in the 1/3 of the document

$I(S_i)$ = $I(S_i)$ + 1

if the sentence contains Concept NAICS or if the sentence contains organization NE

$I(S_i)$ = $I(S_i)$ + 1

if sentence contains URL or email or money

$I(S_i)$ = 1/4

In the next section, we will conduct the experiments with the methods described in this chapter.

CHAPTER 5

EXPERIMENTS

In this chapter, we will describe our experiments on classifying a set of CFT documents. We will test the impact of incorporating the three factors described in the last chapter to see if and how much they are effective in increasing the accuracy of traditional classification algorithms (Naïve Bayes and SVM).

5.1 Test collection

In order to test our proposed methods in chapter 4, we use the call for tender test collection compiled by Francois Paradis for his experiments(Paradis, 2005). The test collection of call for tender documents is created by downloading the XML daily synopsis from the FedBizOpps Web site (tenders solicited by American government agencies, available at http://www.fedbizopps.gov/). The XML documents have the same contents as the HTML documents found on the same site. The period downloaded ranged from September 2000 to October 2003. This test collection has only one document per tender solicitation (in some other cases, a call for tender can contain both solicitations and amendments). There are 21,945 documents (72 MB) in the test collection. They were split 60% for training, and 40% for testing. The classification of FBO documents consists of classifying documents into categories determined by NAICS code (North American Industry Classification System). All these documents have been assigned a NAICS code

manually. The task of automatic classification is to learn the classification model from the training set, and to assign a NAICS code to each of the test document automatically.

Below is an example of FBO CFT document:

<PRESOL>

<DATE> 0521

<YEAR> 01

<CLASSCOD> 75

<NAICS> 424120

<OFFADD> Office of Environmental Studies; 1323 Y Street, Washington, DC 22030

<SUBJECT> Office supplies and devices

<SOLNBR> N00140-04-Q-4555

<ARCHDATE> 07131999
<CONTACT> Mary Ann Deal, Contract Specialist

<DESC> The office of Environmental Studies intends to procure printer toner cartridges and supplies for the Naval Inventory Control Point in Mechanicsburg, PA.

Request for Quotation (RFQ) N00140-04-Q-4555 contemplates an indefinite delivery type price order. This is a combined synopsis/solicitation for commercial items prepared in accordance with the format in FAR Subpart 13.5, Test Program for certain Commercial Items, as supplemented with additional information included in this notice. This announcement constitutes the only solicitation; proposals are being requested, and a written solicitation will not be issued. This is a 100% Total Small Business Set-Aside. etc.

<URL> http://www.oes.gov

<EMAIL>

<ADDRESS> johndoe@usa.gov

<SETASIDE> Total Small Disadvantage Business

<POPZIP> 22030

<POPCOUNTRY> US

</PRESOL>

Figure 5.1: Sample CFT on FBO

As shown above, an FBO call for tender document includes some meta-data such as the classification codes (*<NAICS> 424120*), date of publication (*<DATE> 0521 <YEAR> 01*), the contracting authority (*Office of Environmental Studies*), etc. The body of the document is composed of the subject line (<SUBJECT>) and the description (<DESC>). Only these last two fields will be used for classification. As we explained earlier, only part of the body (the first paragraph of <DESC> in the above example) indicates the subject of the call and is considered to be relevant for classification. The rest concerns dates, the address of the contracting authority, amount of the contract and other standard submission procedures.

The documents have been classified with two classification codes, Product/Service Classification (PSC) Code (*<CLASSCOD>*) and NAICS (*<NAICS>*). However, we will only use NAICS in our study, as both codes play similar roles. The NAICS codes were extracted from the text description and tagged in XML documents.

5.2 NAICS Classification System

The North American Industry Classification System (NAICS) is a classification system originally developed using a production oriented conceptual framework, jointly by US, Canada and Mexico. Its main purpose is for business and government to group establishments into industries based on the activity in which they are primarily engaged. Establishments using similar raw material inputs, similar capital equipment, and similar labor are classified into the same industry. In other words, establishments that do similar things in similar ways are classified together. For example, the manufacturing plants that make pen and mechanical pencil are grouped together in the "Pen and Mechanical Pencil Manufacturing" industry, represented by the code 339941.

Below are some other sample codes of NAICS:

Sector	33	Manufacturing
Subsector	339	Miscellaneous Manufacturing
Industry Group	3399	Other Miscellaneous Manufacturing
Industry	33994	Office Supplies (except Paper) Manufacturing
U. S. Industry	339941	Pen and Mechanical Pencil Manufacturing

Table 5.1: Sample NAICS Category

Another important characteristic of the NAICS codes is that it is hierarchical. As shown in the table above, every digit of a six-digit code corresponds to a level of the hierarchy. For example, for US industry code 339941 (Pen and Mechanical Pencil Manufacturing), the sub sector code is 339 (Miscellaneous Manufacturing) and the Sector code is 33 (Manufacturing). Each of the three participating countries, the U.S., Canada and Mexico, has their own version of the standard, which mostly differ at the level of industry codes (5th or 6th digit). It is very difficult to automatically classify documents into fine-grained classes. This would be possible for larger-grained classes. Therefore, we reduce the problem by considering only the first three digits of NAICS categories, which correspond to sub-sectors. Therefore, only 92 categories are used for the classification of call for tender documents.

5.3 Distribution of the documents

Documents are distributed unevenly in classes. This corresponds to the real situation: there are much more business activities in some sectors than the others. This uneven distribution is kept as it is in our training and test sets. The figure below shows the distribution of documents for NAICS categories. 34% of documents are in the top two categories.

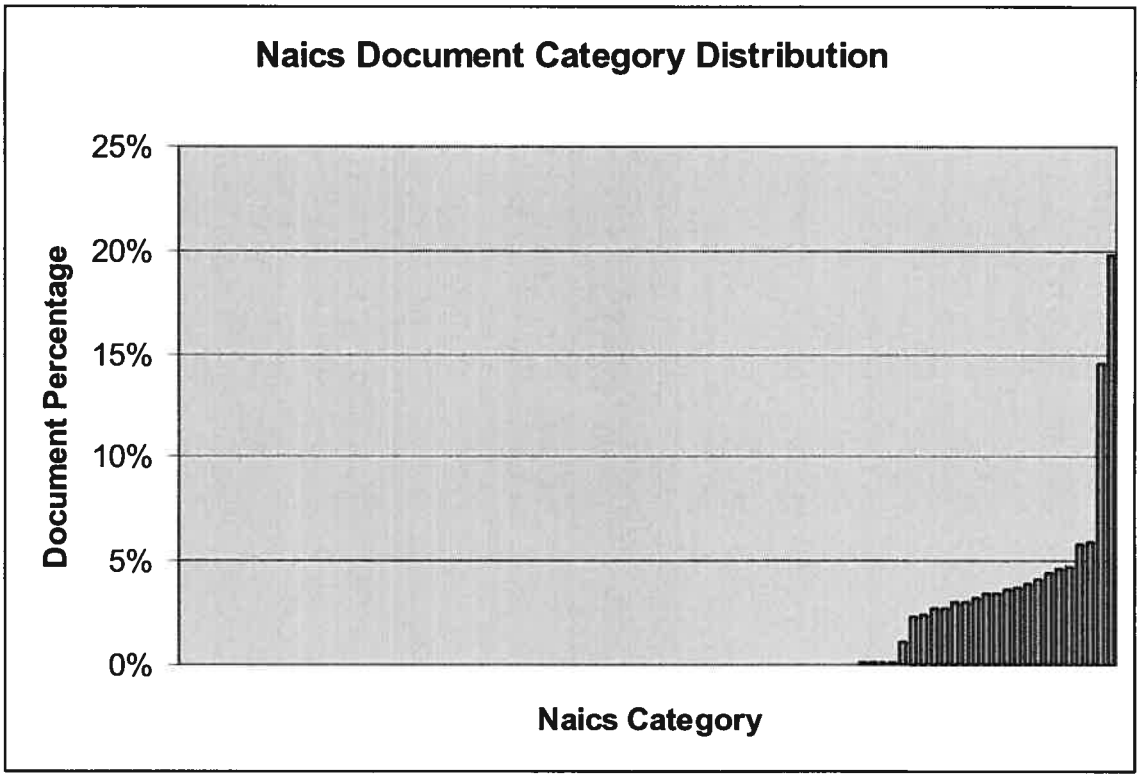

Figure 5.2: Naics Document Category Distribution

5.4 Classifiers

We use two common classifiers in our experiments: Naïve Bayes and SVM. Both classifiers are widely used in text classification. SVM is known to produce better than the average text classification result (Joachims, 1998). Naïve Bayes is a simple classifier.

Although it usually produces a lower accuracy than SVM, it is also widely used for its high efficiency and reasonable accuracy. Therefore it is appropriate to choose these two classifiers for our study. The baseline for both classifiers is the classifier trained and tested on the unfiltered documents. In the baseline classifiers, no special treatment is performed on the documents, except the standard stemming and stopword removing.

On the baseline classifiers, we also applied a classical feature selection method – InfoGain. The experiment parameters have been optimized for this type of call for tenders collection:

- For both Naïve Bayes and SVM, the 8,000 top terms (features) were selected according to their InfoGain score - this produced the best accuracy.
- Also, the following thresholds were applied: a rank cut (rcut) of 1 (i.e. to select the best ranked class for each document) and a score-based cut (scut) learnt for each category after cross-sampling 50% of the training set over 10 iterations. At the end of cross-validation, each category is associated with the best score threshold to determine if document should be assigned to a class according to the score of the class.

The setup of the InfoGain experiment and its explanation can be referred to (Paradis, 2005).

The *rainbow* classification software was used to perform our experiments. *Rainbow* is a program developed by Andrew McCallum that performs statistical text classification (McCallum, 1996). It can perform text classification with various methods such as SVM, Naïve Bayes, KNN, etc. It also allows users to modify the term weights in

the indexing step of the classification process, which is useful for our study. This software is used for both Naïve Bayes and SVM classification of CFT documents.

5.5 Evaluation of Classification Experiment

Similar to the information retrieval, precision and recall are the performance measures of the classification experiment. Given a document as the input of the classification and a list of the categories as output, the precision and recall are defined to be:

$$\text{Precision} = \frac{\text{categories found and correct}}{\text{total categories found}}$$

$$\text{Recall} = \frac{\text{categories found and correct}}{\text{total categories correct}}$$

We can evaluate the performance of a binary classifier using a contingency matrix.

	Correct	Incorrect
Assigned class	**a**	**b**
Unassigned class	**c**	**d**

Table 5.2: Contingency matrix

where **a** is the number of assigned correct cases

b is the number of assigned incorrect cases

c is the number of unassigned correct cases

d is the number of unassigned incorrect cases

The recall and precision can be calculated as the following:

r (recall) = a / (a+c) if a + c > 0 otherwise r = 1

p (precision)= a / (a+b) if a + b > 0 otherwise p = 1

There is a tradeoff between precision and recall. One of the most common methods of evaluation that combines precision and recall is F-measure:

$$F_\beta(r,p) = \frac{(\beta^2+1)pr}{\beta^2 p + r}$$

where p = precision

r = recall

β = weighting parameter that favors either precision or recall

if β is one, then precision and recall are weighted equally. F-measure becomes the F_1-measure:

$$F_1(r,p) = \frac{2pr}{p+r}$$

In order to evaluate the performance of binary classifiers like SVM and Naïve Bayes, we can use two averaging methods over all the classification results: micro-averaged F1 (Micro-F1) and macro-averaged F1 (Macro-F1) metric. Macro-F1 uses one contingency matrix per category. The local measures are computed first for each category and then averaged over categories. Micro-F1 merges the contingency matrix of each

category into one contingency matrix. **a, b, c, d** are the sum of the corresponding cells in the local table. Macro-F1 gives equal weight to each F1 score without regarding to how common the category is. The Micro-F1 metric gives equal weight to all classifications, so that F1 scores of more common classes influence the metric more than F1 scores of less common classes.

5.6 Baseline classifiers

	Base line: Naïve Bayes	Base line: SVM
miF1	0.51196	0.63495
maF1	0.10764	0.36927

Table 5.3: Baseline Classifiers Performance

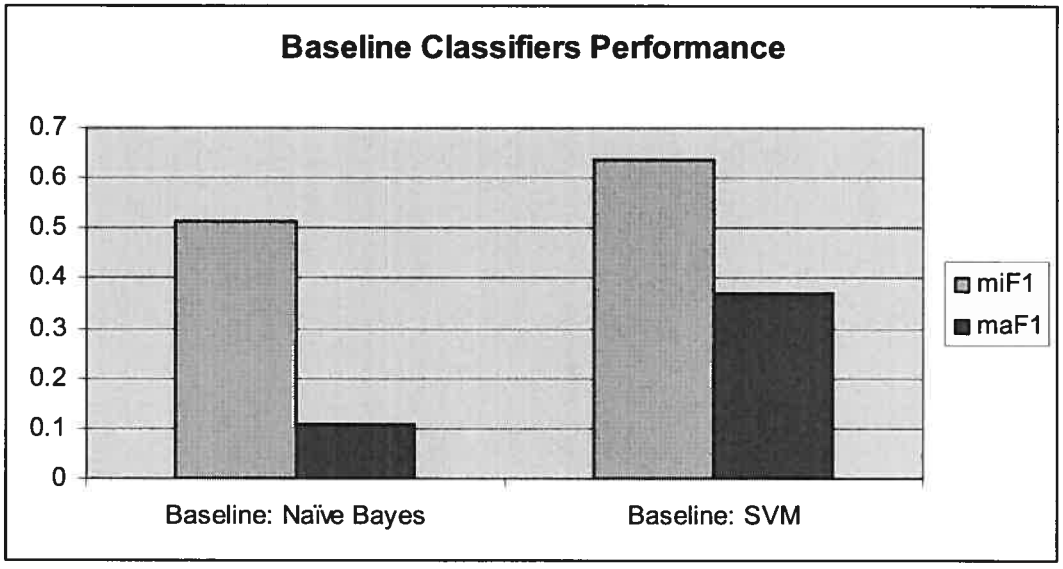

Figure 5.3: Baseline Classifiers Performance

For the baseline classifiers without feature selection or term weightng method, the performance of SVM classifier is much better than the performance of Naïve Bayes for this type of call for tenders classification. The micro-F1 of SVM is 0.63495 compared to

0.51196 for Naïve Bayes. It confirms previous studies that SVM performance text classification very well.

5.7 Text Classification Using Position Information

5.7.1 Sentence Filtering By Position Information:

We know the first 1/3 sentences are the most relevant according the experiments described in section 4.1.1. This features selection experiment using position information consists of keeping only the first 1/3 sentences of each document and filtering out the rest of the document. Then we run Naive Bayes classifier or SVM classifier using the filtered corpus. The table and the figure below show the results obtained for Naïve Bayes.

	Baseline	First 1/3	InfoGain
Micro-F1	0.51196	0.52985 (+3.49%)	0.52711 (+2.96%)
Macro-F1	0.10764	0.13117 (+21.86%)	0.24055 (+123.48%)

Table 5.4: Position feature selection (Naive Bayes)

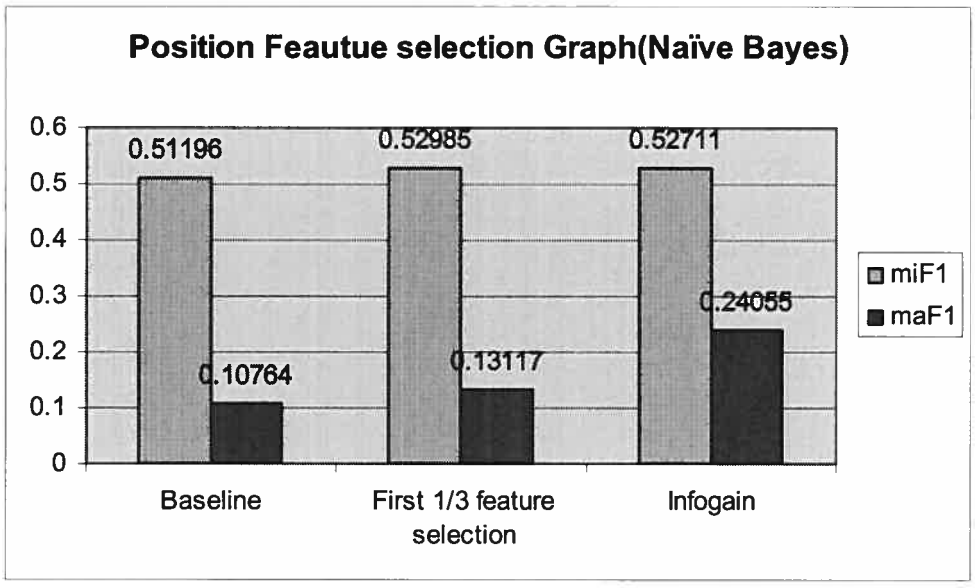

Figure 5.4: Feature selection by position (Naive Bayes)

We see that the classification result is better than the baseline: a Micro-F1 of 0.52985 (+3.49%) and a Macro-F1 of 0.13117 (+21.86%). The Micro-F1 of sentence filtering method is slightly better than the Micro-F1 with InfoGain method with no position information: 0.52711 (+2.96%).

We perform the same test with SVM. The results are reported in the following table and figure.

	Baseline	First 1/3 feature selection	InfoGain
Micro-F1	0.63495	0.61881 (-2.54%)	0.63461 (-0.053%)
Macro-F1	0.36927	0.37635 (+1.92%)	0.36908 (-0.051%)

Table 5.5: Position feature selection (SVM)

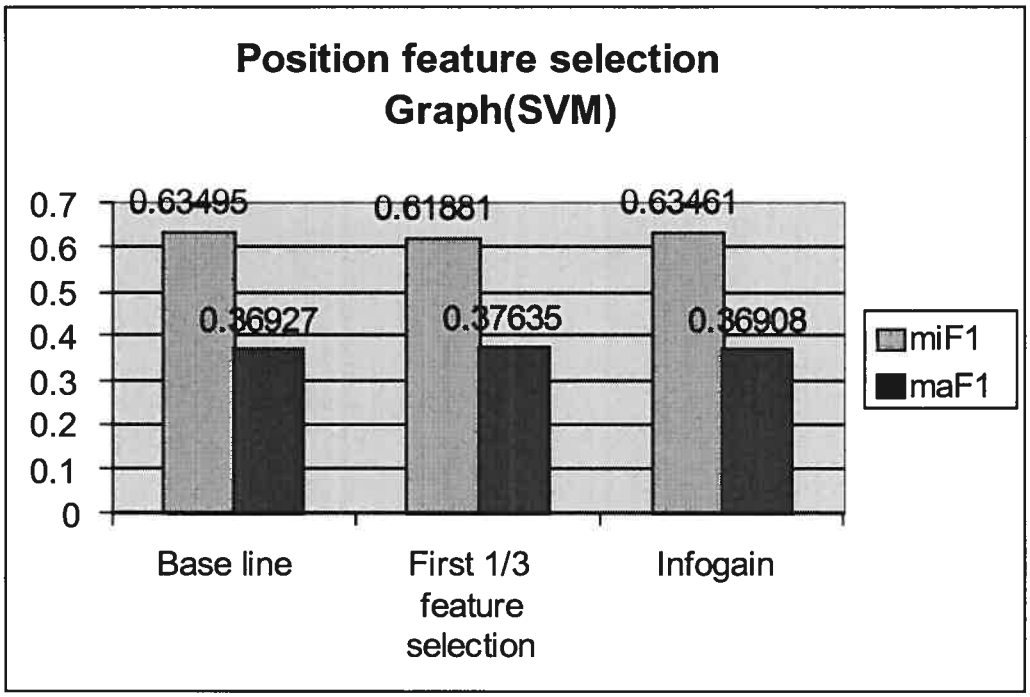

Figure 5.5: Position feature selection (SVM)

With the SVM classifier, we see that the classification result is not improved with respect to the baseline: a Micro-F1 0.61881 (-2.54%) and a Macro-F1 of 0.37635 (+1.92%). The micro-F1 of position feature selection method is worse than the micro-F1 with InfoGain feature selection of 8000 terms. However, the macro-F1 of position feature selection method is slightly better.

As we can see, the sentence filtering with positional information method doesn't help to improve the classification result. A possible reason is that the filtering process is too strong: a sentence is either considered to be important or not. The non-important sentences then are not considered at all during the classification process. As we can see from the earlier experiments, position cannot accurately determine all and only the important sentences. Some important sentences will be tagged "non-important" and some

other non important sentences tagged "important". Therefore, we cannot rely heavily on the result of this identification for selection of important features.

An alternative that we will consider is to assign weights to terms according to the position of the sentence in which they appear. This method is less strict than filtering: even if a sentence is not tagged "important", the terms appearing in it will still be considered to some extent in the classification process.

5.7.2 Term Weighting Using Position Information

The original term weight is the term frequency of occurrence in the document. We want to increase the weight of the terms appearing in the important sentences. We experiment with two methods: The first one consists of increasing the term frequency of first 1/3 sentences of the document by multiplying it by a boosting factor 8. The second methods increases the frequency of first 1/3 sentences by multiplying by 8 and the frequency of sentences between 1/3 and 2/3 of the document by multiplying by 4, and the frequency of features in the final 1/3 sentences of the document remain unchanged. This second method assigns the term weights according the relative importance of the sentence position in the document. The boosting factor is selected empirically after experiments with several different values (please see figure 5.7).

Then the resulting term weights are passed to Naive Bayes and SVM classifiers. The table and figure below show the results with Naïve Bayes.

	Baseline	First 1/3	First1/3 (8x),second 1/3 (4x),last1/3 (1x)	InfoGain
Micro-F1	0.51196	0.60014 (+17.22%)	0.60442 (+18.06%)	0.52711 (+2.96%)
Macro-F1	0.10764	0.26329 (+144.60%)	0.2834 (+164.29%)	0.24055 (+123.48%)

Table 5.6: Term Weighting by Position (Naive Bayes)

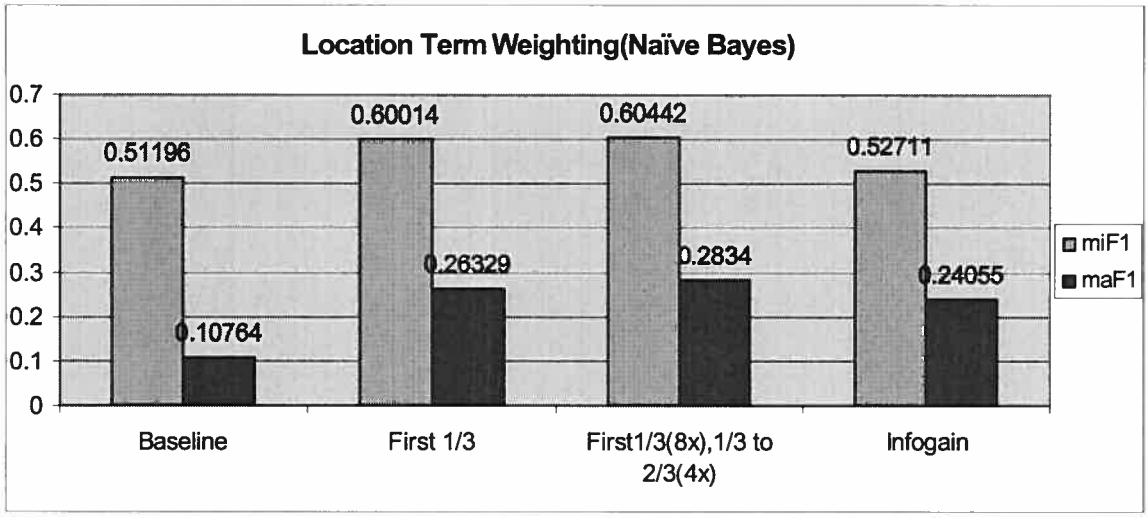

Figure 5.6: Term Weighting by Position (Naive Bayes)

For Naïve Bayes classifier, the graph shows a very significant improvement when we increase only the frequencies first 1/3 of the document in the Naïve Bayes Classification experiment. The Micro-F1 has increased to 0.60014(+17.22%), which is much better than the baseline. With the second method that assigns the term weights according the relative position of the sentences in the document (first 1/3: 8x, second 1/3:

4x , final 1/3: 1x), the result is even better than the first method. The Micro-F1 is 0.60442 (+18.06%) and macro-F1 is 0.2834 (+164.29%).

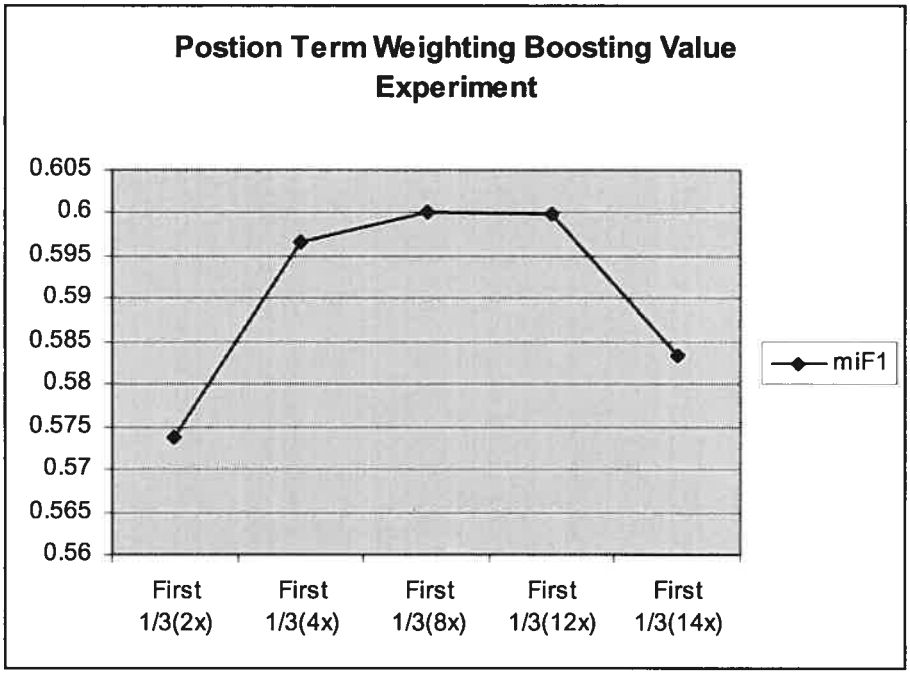

Figure 5.7: Position Term Weighting Boosting Value Experiment (Naïve Bayes)

Figure 5.6 shows that with the method that increases only the frequencies first 1/3 of the document, the best classification result comes if the boosting factor is 8.

The same boosting factors have been used for term weighting in SVM. The results are shown below.

	Baseline	First 1/3	First1/3(8x), second 1/3 (4x)	InfoGain
Micro-F1	0.63495	0.62173	0.63958 (+0.729%)	0.63461 (-0.053%)
Macro-F1	0.36927	0.38114	0.4021 (+8.89%)	0.36908 (-0.051%)

Table 5.7: Term Weighting by Position (SVM)

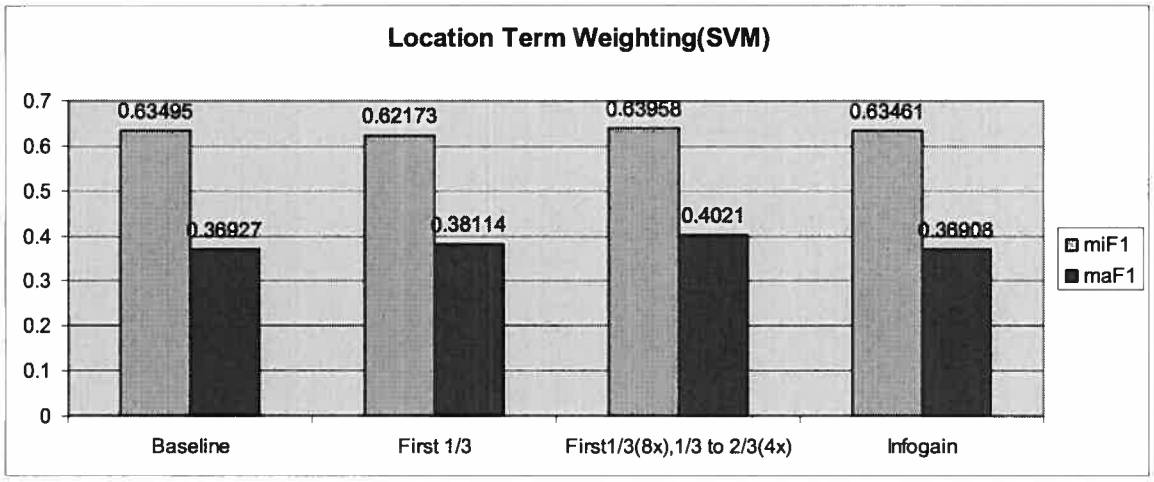

Figure 5.8: Term Weighting by Position (SVM)

For SVM classifier, the results obtained by increasing term weights for the features in sentences in first 1/3 are worse than the baseline. With the second method that assigns the term weights according the relative position of the sentences in the document (first 1/3: 8x, second 1/3: 4x, final 1/3: 1x), the classification result is only slightly better than the baseline with the micro-F1 of 0.63958 (+0.729%). This result is consistent with the previous studies which have found that additional feature selection can hardly improve the quality of the classification for SVM classifier. The reason is because the SVM method already has the ability of taking into account the importance of feature in its algorithm.

5.8 Text Classification Using Concepts

5.8.1 Sentence Filtering Using Concepts

This method consists of filtering out all the sentences that don't contain at least one concept that has IR score over 0.8. Then we run Naïve Bayes classifier or SVM

classifier using the filtered corpus. As shown in figure 5.11, the threshold value 0.8 is chosen as the best empirical value for Micro-F1 among a few threshold values tested.

For the Naïve Bayes classifier, the concept sentence filtering method show an improvement of classification result over the baseline we have established. The improvement of Micro-F1 is +2.67% and the improvement of Macro-F1 is +25.2%. The classification result improvements are not as significant as the improvement made by InfoGain features selection method. It's probably because of the inaccuracy of the NConcept Extractor. Many irrelevant concepts remain even after the concepts go through information retrieval on NAICS definition and stoplist. Also compared to other methods that retain the whole document, filtering out wrong concepts can decrease the precision measure, which affects the overall F1 measure.

	Baseline	Concept Feature Selection	Concept Term Weighting	InfoGain
Micro-F1	0.51196	0.52563 (+2.67%)	0.55901 (+9.19%)	0.52711 (+2.96%)
Macro-F1	0.10764	0.13476 (+25.2%)	0.20778 (+93.03%)	0.24055 (+123.48%)

Table 5.8: Concept feature selection and term weighting (Naïve Bayes)

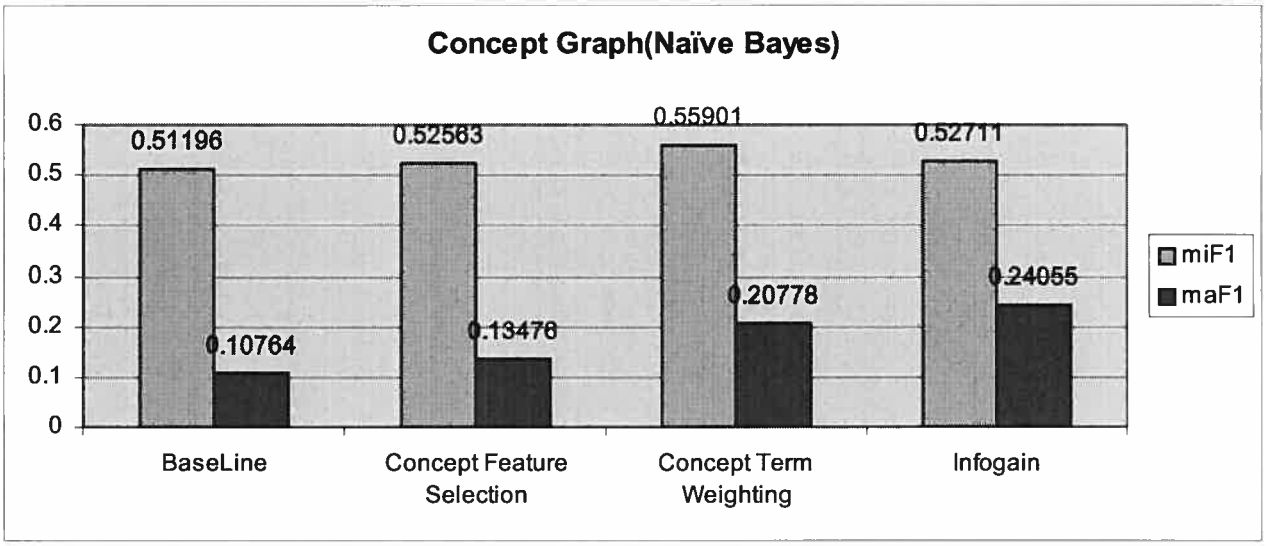

Figure 5.9: Concept feature selection and term weighting (Naïve Bayes)

	Baseline	Concept Feature Selection	Concept Feature Weighting	InfoGain
Micro-F1	0.63495	0.59323 (-6.57%)	0.60734 (-4.35%)	0.63461 (-0.05%)
Macro-F1	0.36927	0.35008 (-5.2%)	0.35391 (-4.16%)	0.36908 (-0.05%)

Table 5.9: Concept feature selection and term weighting (SVM)

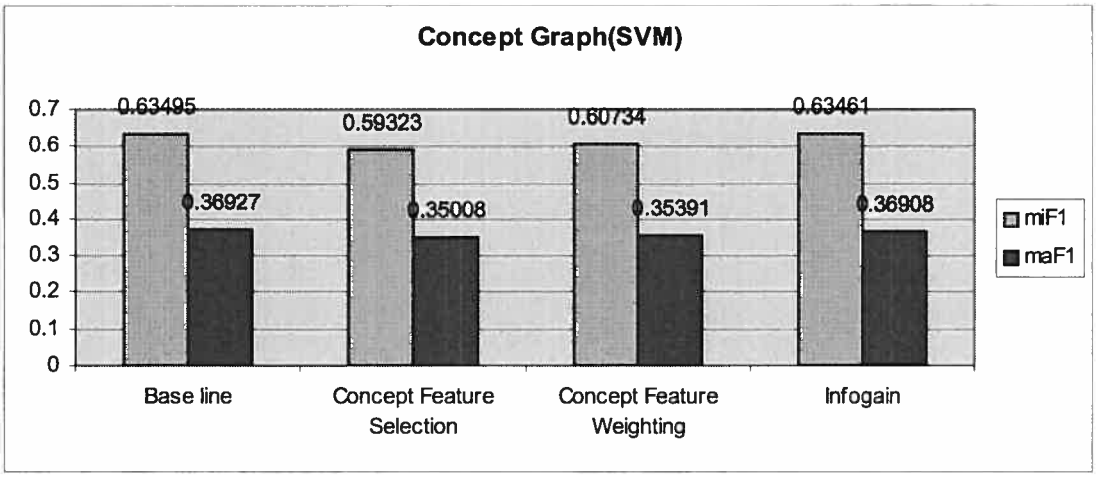

Figure 5.10: Concept feature selection and term weighting (SVM)

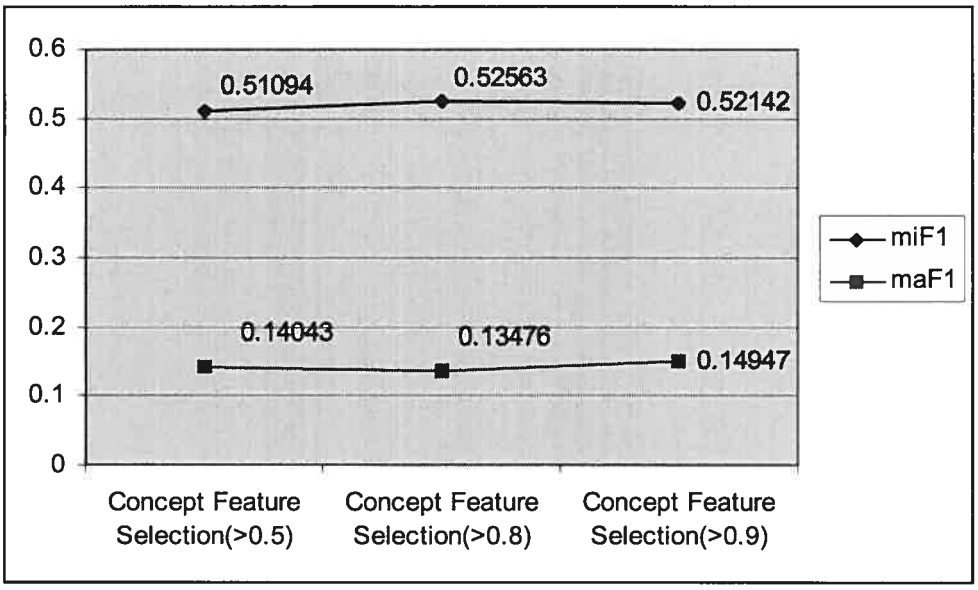

Figure 5.11: Concept Feature selection threshold value experiment (Naïve Bayes)

For the SVM classifier, the concept features selection method shows a classification result worse than the method without it. It performs even worse than the InfoGain method. This is possibly because SVM has the ability to generate well in dimensional feature spaces. Since SVMs use overfitting protection that does not depend on the number of features. Generally the performance improvement of SVM classifier using feature selection method hasn't been proved to be significant. Also, it's probably because of the inaccuracy of the NConcept Extractor, which decreases the precision measure, so affects the overall F1 measure.

5.8.2 Term Weighting According to Concepts

Experiment consists of increasing the term frequencies of sentences containing filtered concepts. We tested several combination of increasing the term weights of concepts with IR score. Figure 5.12 shows that the best concept boosting factor value

comes when the term weights of features in sentences containing concept with IR score over 0.8 are multiplied by 4. The term weights of features in sentences containing concept with IR score between 0.5 and 0.8 is multiplied by 2. Sentences that don't contain concepts or containing concepts with IR score less than 0.5 keep the same term weights.

With the term weighting method, the Naïve Bayes classification result shows a better improvement compared to the sentence filtering method with concept extraction: Micro-F1 improves to 0.55901 (+9.19%) compared to 0.52563 (+2.67%) with the sentence filtering feature selection method. The Micro-F1 is also better than the one with InfoGain feature selection method: 0.52711 (+2.96%). But the Macro-F1 (0.20778) is worse than the one with the InfoGain feature selection method (0.24055).

For the SVM classifier, the classification result of term weighting method shows a result worse than the baseline. Both Micro-F1 and Macro-F1 measures are worse than the baseline result, -4.35% and -4.16% respectively.

	Term Weighting(6,2,1) >=0..8-->x6 >=0.5 and <0.8→ x2 <0.5 →x1	Term Weighting(5,2,1) >=0..8-->x5 >=0.5 and <0.8→ x2 <0.5 →x1	Term Weighting(4,2,1) >=0..8-->x4 >=0.5 and <0.8→ x2 <0.5 →x1	Term Weighting(3,2,1) >=0..8-->x3 >=0.5 and <0.8→ x2 <0.5 →x1
Micro-F1	0.5382	0.5373	0.55901	0.51283
Macro-F1	0.1432	0.18733	0.20778	0.18643

Table 5.10: Term Weighting by Concepts (Naive Bayes)

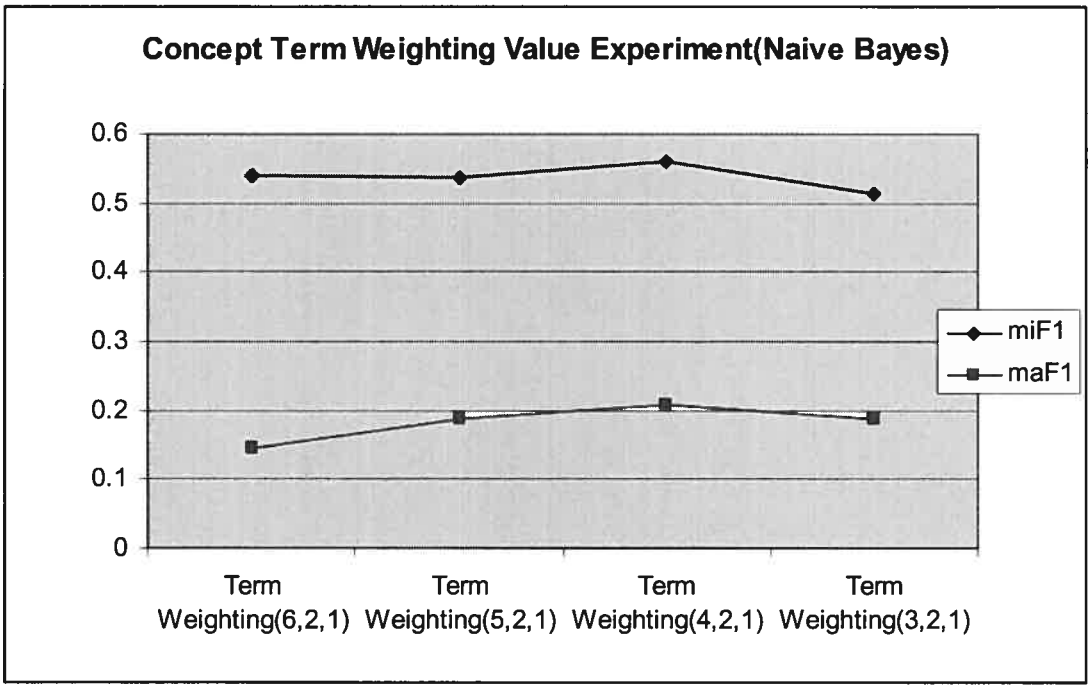

Figure 5.12: Concept Term Weighting by Concepts (Naive Bayes)

5.9 Text Classification using Named Entities

5.9.1 Sentence Filtering Using NE

From the experiments described in section 4.1.3 we know that money, URL and email negative indicators of relevance and organization can be seen as positive indicator. The other types of named entities are more ambiguous with respect to their indication of important sentences. The NE feature selection method consists of filtering out all the sentences that contain money, URL and email.

We see a slight improvement of Micro-F1 over the baseline (+1.54%). The improvement of Macro-F1 over the baseline is much larger (+127.11%). Compared to the

result obtained from the InfoGain feature selection method, the classification result is similar: Micro-F1 of 0.51982 vs. 0.52711 and Macro-F1 of 0.24446 vs. 0.24055.

	Baseline	NE Feature Selection	NE Term Weighting	InfoGain
Micro-F1	0.51196	0.51982 (+1.54%)	0.5704 (+11.41%)	0.52711 (+2.96%)
Macro-F1	0.10764	0.24446 (+127.11%)	0.24062 (+123.54%)	0.24055 (+123.48%)

Table 5.11: NE feature selection and term weighting experiments (Naive Bayes)

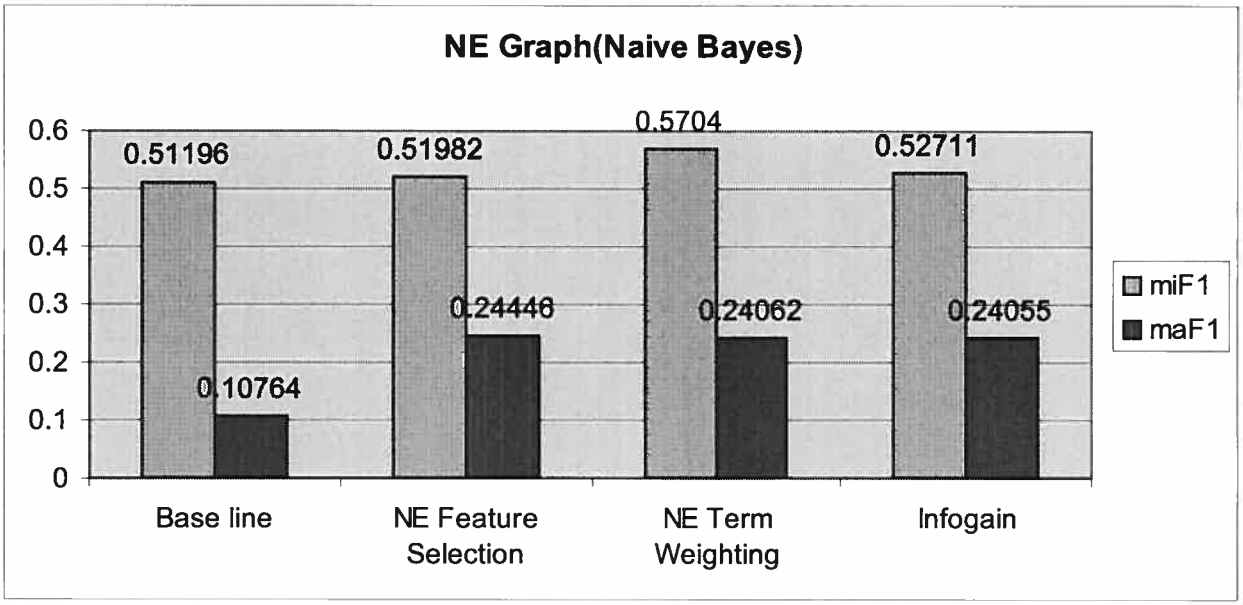

Figure 5.13: NE feature selection and term weighting experiments (Naive Bayes)

	Base line	NE Feature Selection	NE Term Weighting	InfoGain
Micro-F1	0.63495	0.63018 (-0.75%)	0.62649 (-1.33%)	0.63461 (-0.05%)
Macro-F1	0.36927	0.37637 (+1.92%)	0.38007 (+2.92%)	0.36908 (-0.05%)

Table 5.12: NE feature selection and term weighting experiments (SVM)

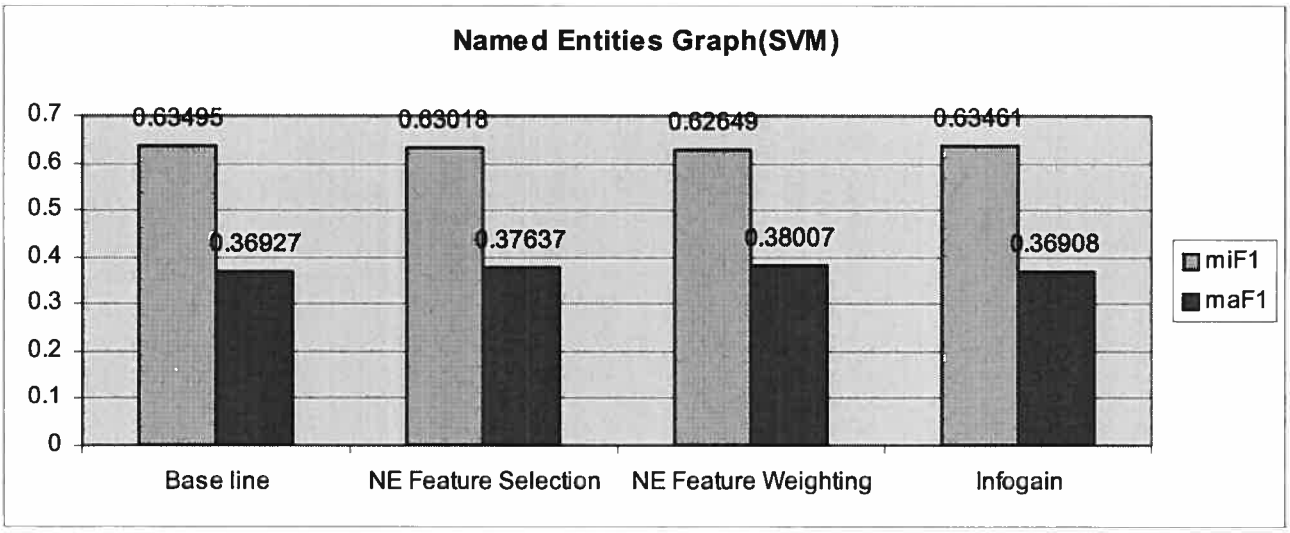

Figure 5.14: NE feature selection and term weighting experiments (SVM)

For the SVM classification experiment, not much improvement is observed with the NE feature selection method. The Micro-F1 of the NE feature selection experiment is slightly worse than the baseline, 0.62649 (-1.33%) and the Macro-F1 measure is slightly better than the baseline (+1.92%).

5.9.2 Term Weighting According to Named Entities

	No Neg. NE(x2), Organization(x4)	No Neg. NE(x4), Organization(x8)	No Neg. NE(x6), Organization(x12)	No Neg. NE(x8), Organization(x16)
Micro-F1	0.5573	0.5704	0.56086	0.55802
Macro-F1	0.20346	0.24062	0.22219	0.20232

Table 5.13: Term Weighting by NE with different Boosting factors (Naïve Bayes)

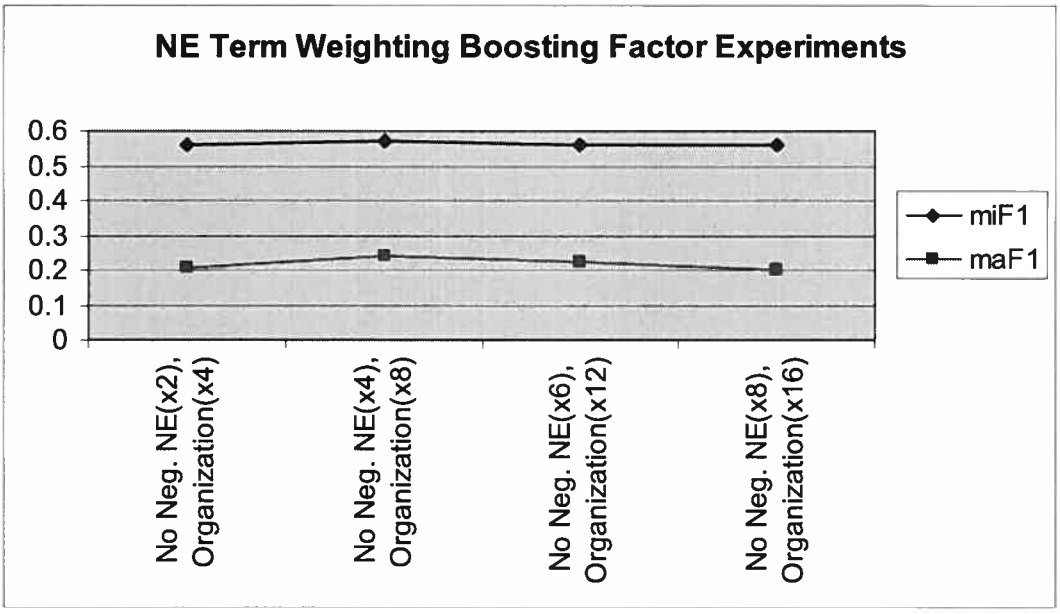

Figure 5.15: Term Weighting by NE with different Boosting factors (Naïve Bayes)

This experiment consists of increasing the weight of the term frequencies of the sentences containing the organization and decreasing the term frequencies of the all the sentences that contain money, URL and email. The figure 5.15 shows the best experimental result comes with the following weighting method: the term weights of

sentences that don't contain negative NE mentioned above are increased by 4, the term weights of the sentences with organization are increased by 4x2=8, the term weights of sentences that contain URL and email will stay the same.

Naïve Bayes classification result shows a significant improvement of Micro-F1 (+11.41%) and Macro-F1 (+123.54%) over the baseline. The result is also better than the result obtained from InfoGain features selection method. Compared to the sentence filtering method, the term weight method result has better Micro-F1 (0.5706 vs. 0.51982). However, the Macro-F1 is slightly worse (0.24062 vs. 0.24446).

SVM classification with new term weighting didn't show much improvement. The Micro-F1 of the NE term weighting experiment is slightly worse than the baseline, 0.62649 (-1.33%) and the Macro-F1 measure is slightly better than the baseline, 0.38007 (+2.92%).

5.10 Term Weighting Combining Various Factors

As stated in chapter 4, the positive indicators are organization, concept and 1/3 sentences of each document. The negative indicators are the money, URL and email. In this experiment, we decrease the term frequencies of the all the sentences that contain money, URL and email. We increase the weight of the term frequencies of all the sentences in the first 1/3 of each document. For those sentences containing organization and concepts inside the sentences of 1/3 of the document, we increase more the term frequencies.

The detail of this method is described in section 4.5. One problem is to select a good boosting factor value. As the purpose of this method is to verify whether a combination of different features can help to improve the term weighting method instead

of finding the best combination possible, the boosting factor value is selected after conducting several Naïve Bayes classification experiments with different values. The figure and table below show that for Naive Bayes, the best boosting factor value is 4. Therefore, we also use boosting factor 4 for SVM experiment. It's possible that better combination of those features can be found with machine learning techniques.

	Boosting Factor=2	Boosting Factor=4	Boosting Factor=6
Micro-F1	0.60232	0.62044	0.61413
Macro-F1	0.23322	0.24062	0.35423

Table 5.14: All factor Term Weighting Boosting Factor Experiments(Naïve Bayes)

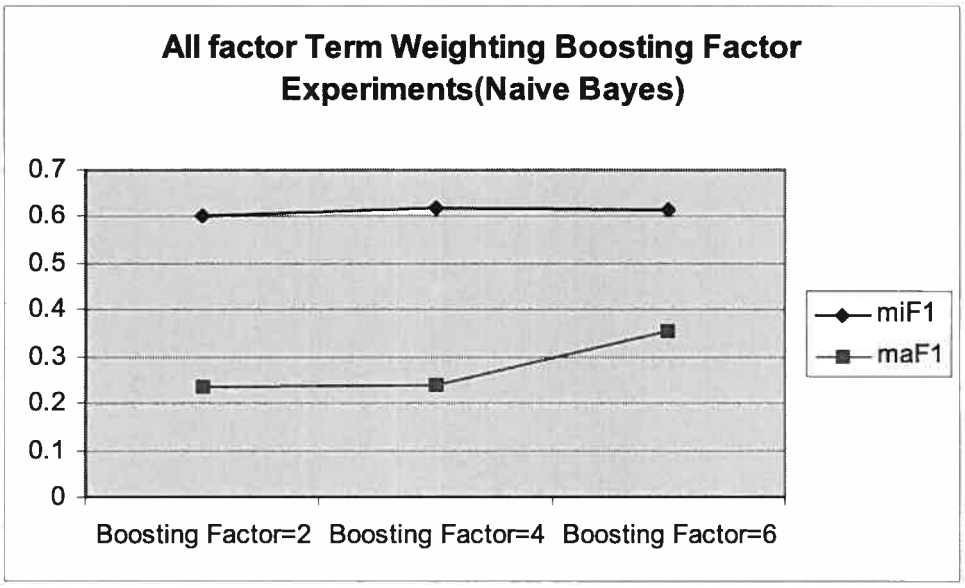

Figure 5.16: All factor Term Weighting Boosting Factor Experiments(Naïve Bayes)

5.10.1 Naïve Bayes Classification With Combined Factors

The classification result of the term weighting methods is much higher than the base line: micro-F1 of 0.62044 (+21.19%) and macro-F1 of 0.29788 (+176.74%).

	Base line	All factors Term Weighting	InfoGain
Micro-F1	0.51196	0.62044 (+21.19%)	0.52711 (+2.96%)
Macro-F1	0.10764	0.29788 (+176.74%)	0.24055 (+123.48%)

Table 5.15: Term weighting incorporating all factors (Naive Bayes)

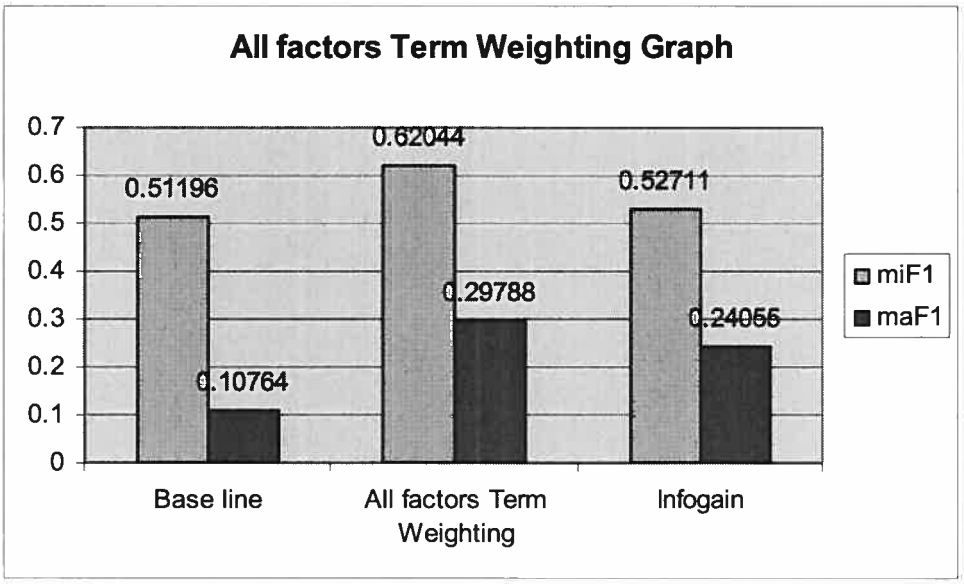

Figure 5.17: Term weighting incorporating all factors (Naive Bayes)

5.10.2 SVM with combined factors

This SVM experiment uses the same algorithm as the one for Naïve Bayes. The classification result of the term weighting methods is slightly higher than the baseline:

Micro-F1 of 0.64392 (+1.41%) and Macro-F1 of 0.4061 (+9.97%). Even though the improvements are small, we can still conclude that the consideration of the new factors in term weighting can help improve SVM accuracy. This result is very encouraging because it has been found in the previous studies that it is very difficult to improve SVM. So even a small improvement can be considered as a success.

	Baseline	All factors Term Weighting	InfoGain
Micro-F1	0.63495	0.64392(+1.41%)	0.63461(-0.05%)
Macro-F1	0.36927	0.4061(+9.97%)	0.36908(-0.05%)

Table 5.16: Term weighting with all factors (SVM)

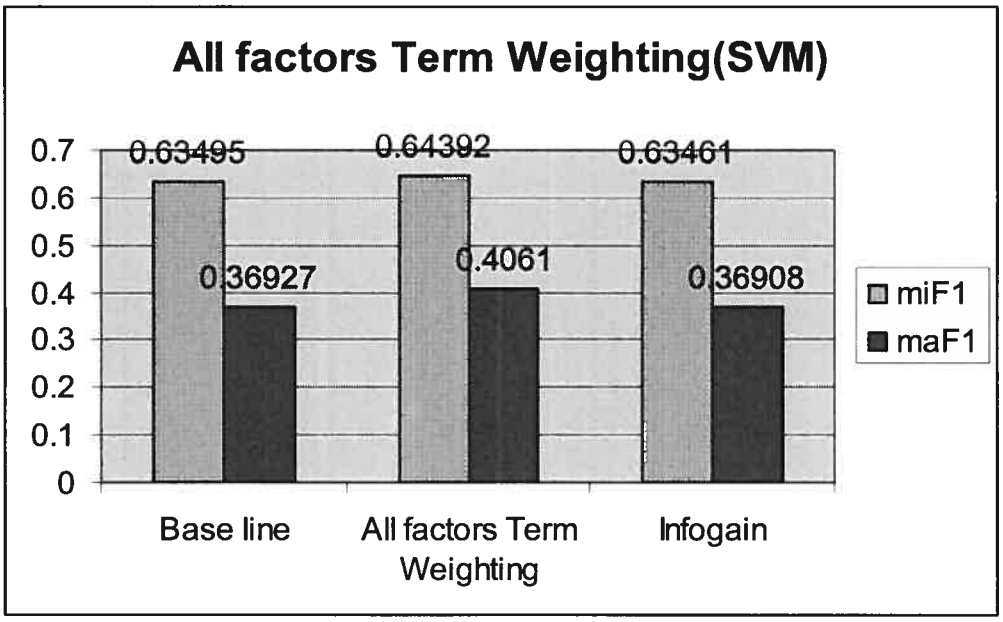

Figure 5.18: Term weighting with all factors (SVM)

5.11 Overall Comparison of Classification Results

The figures and tables below present the classification results of term weighting, sentence filtering and InfoGain methods with all the factors for both SVM and Naïve Bayes classifiers respectively. Also for both SVM and Naïve Bayes classifiers, the average Micro-F1 and Macro-F1 values of term weighting and sentence filtering methods are also presented.

5.11.1 Naïve Bayes Classification Result:

Method	macro-F1	micro-F1
Baseline	0.10764	0.51196
Feature Selection with InfoGain	0.24055	0.52711
Sentence Filtering by position: First1/3	0.13117	0.52985
Term Weighting by position: First1/3	0.26329	0.60014
Term Weighting by position: First1/3 (8x), second 1/3 (4x), Final 1/3(1x)	0.2834	0.60442
Sentence Filtering by Concept	0.13476	0.52563
Term Weighting by Concept	0.20778	0.55901
Sentence Filtering by Named Entities	0.24446	0.51982
Term Weighting by Named Entities	0.24062	0.5704
Term Weighting with All factors	0.29788	0.62044

Table 5.17: Naïve Bayes Classification Result

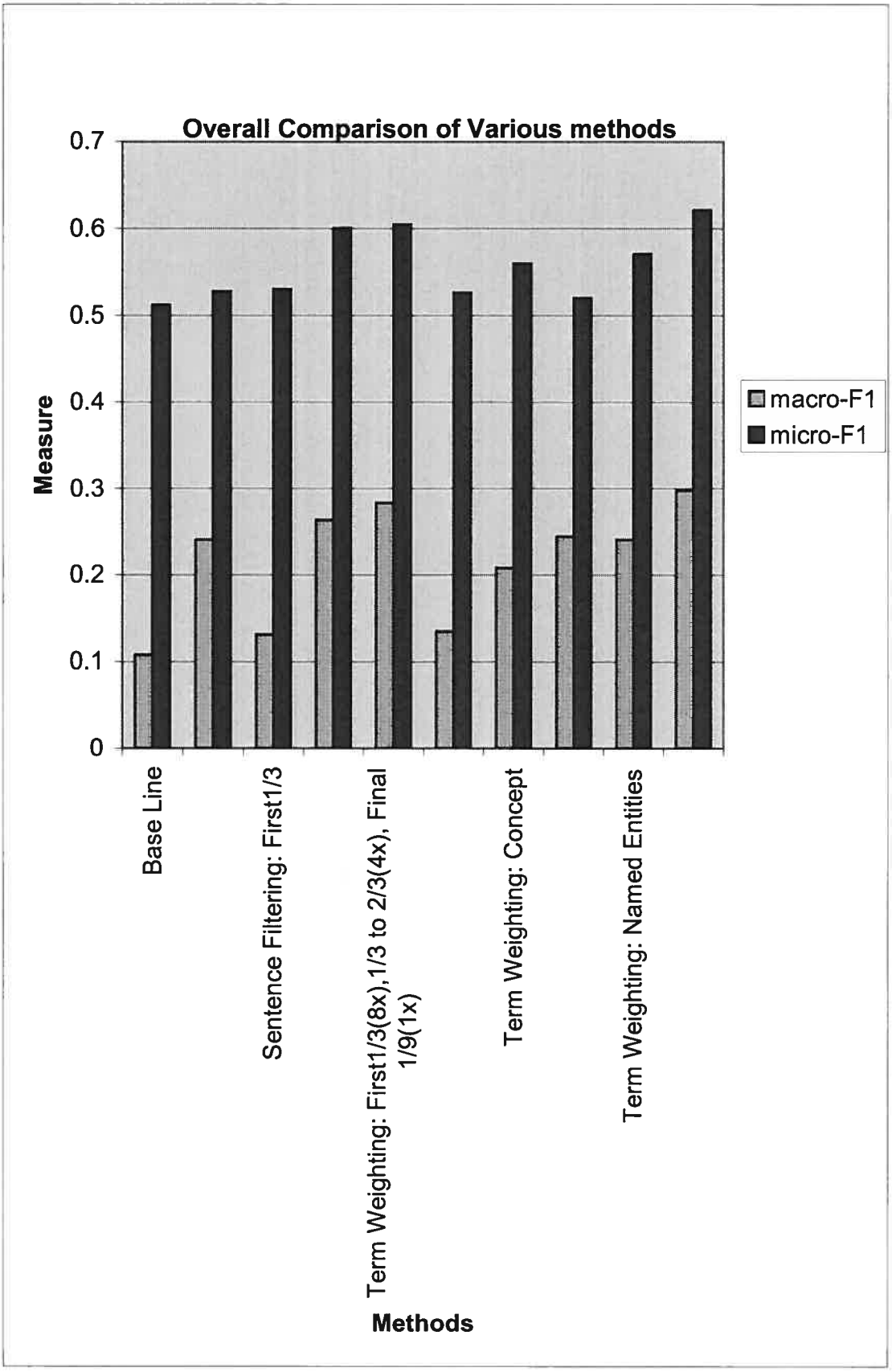
Overall Comparison of Various methods
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Measure
macro-F1
micro-F1
Base Line
Sentence Filtering: First1/3
Term Weighting: First1/3(8x),1/3 to 2/3(4x), Final 1/9(1x)
Term Weighting: Concept
Term Weighting: Named Entities
Methods

Figure 5.19: Naïve Bayes Classification Result

	macro-F1	micro-F1
AVERAGE(Term Weighting):	0.25859	0.59088
AVERAGE(Sentence Filtering):	0.17013	0.5251

Table 5.18: Average Term Weighting and Feature Selection (Naive Bayes)

5.11.2 SVM Classification Result:

Method	macro-F1	micro-F1
Baseline	0.36927	0.63495
Feature Selection with InfoGain	0.36908	0.63461
Feature Selection by position: First 1/3	0.37635	0.61881
Term Weighting by position: First1/3	0.38114	0.62173
Term Weighting by position: First1/3 (8x), second1/3 (4x), Final 1/3(1x)	0.4021	0.63958
Feature Selection by Concept	0.35008	0.59323
Term Weighting by Concept	0.35391	0.60734
Feature Selection by Named Entities	0.37637	0.63018
Term Weighting by Named Entities	0.38007	0.62649
Term Weighting: All factors	0.4061	0.64392

Table 5.19: SVM Classification result

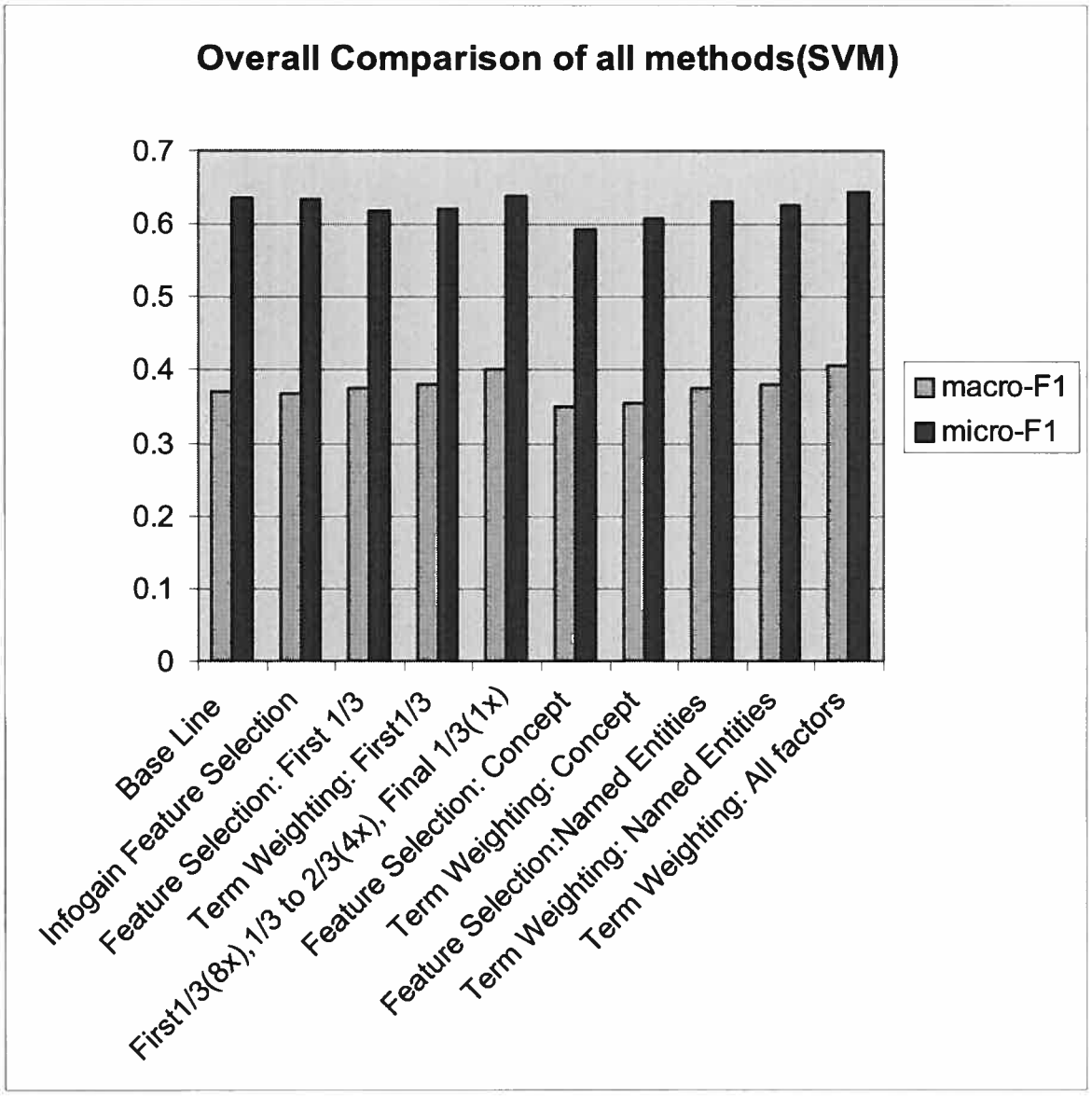

Figure 5.20: SVM Classification Result

	macro-F1	micro-F1
AVERAGE(Term Weighting):	0.38466	0.62781
AVERAGE(Feature Selection):	0.3676	0.61407

Table 5.20: Average Term Weighting and Feature Selection (SVM)

5.11.3 Discussion

Sentence Filtering vs Term Weighting

From the tables above, we see that for both SVM and Naive Bayes methods, term weighting method generally gives better result than the feature selection by sentence filtering method for this particular collection of call for tender documents: For Naive Bayes method, the average Micro-F1 of term weighting is 0.59088 and the average micro-F1 of sentence filtering is 0.5251. For SVM, the average micro-F1 of term weighting 0.62781 and the average micro-F1of sentence filtering is 0.61407. A possible reason of these differences is that in text classification, even irrelevant sentences contain less important terms, they are sometimes still useful. In InfoGain feature selection method for example, even if features are ranked low, they still can contain considerable information and are somewhat useful for classification. The feature selection by sentence filtering method eliminates many features and in a strict way. However, those features can still contain valuable information. The loss of those features can hurt the performance of the classifier. Term weighting method, in contrast, retains all the features but with different weights, which can be more appropriate for text classification.

Comparison of Classification Results with Position Information, Concept and NE

In this thesis, three important factors studied are position information, concepts and named entities. Among all those factors, the term weighting method with position information (First1/3 (8x), second 1/3 (4x), Final 1/3 (1x)) seems to produce the best improvement on classification result for Naive Bayes method with Micro-F1 of 0.60442. The position feature selection method also gives the best classification result compare to feature selection with other factors with Micro-F1 of 0.52985. For SVM method, term weighting method with position information (First1/3(8x), second1/3 (4x), Final 1/3(1x)) also gives the best classification result just ahead of NE term weighting with Micro-F1 of 0.63958. This is slightly better than the baseline SVM classifier. One reason of good classification result with position information might be that this method determines the relevant features very well: Most relevant features are in the first 1/3 of the document. Term weighting method and sentence filtering methods with named entities also give good result for both SVM and Naive Bayes classifiers. The reason might be that money, URL and email give very accurate indication of irrelevant part of the document. On the other hand, term weighting method with concepts doesn't perform as well as the other two methods for both SVM and Naive Bayes experiments. It's probably because the concepts extracted by NConcept Extractor contain many irrelevant concepts: Many irrelevant concepts remain even after the concepts go through information retrieval on NAICS definition and stop list.

Multiple Features

All three factors help identify relevant and irrelevant sentences of the document. The term weighting method using positive and negative indicators of those factors has the best classification result with Naïve Bayes classifier (Micro-F1 of 0.64392) and SVM classifier (Micro-F1 of 0.62044) compared to the performance with single factor involved. One possible cause is that by using all three factors, we identify relevant and irrelevant sentences more accurately compared to term weighting method with just one factor.

NB vs. SVM

The classification results show that both feature selection methods and term weighting methods improve the performance of Naive Bayes classifier more than the performance of SVM classifier. Only the term weighting methods that combine all three factors and term weighting with position information improve the micro-F1 measure of SVM classifier. In contrast, all the features selection and term weighting methods used improve the micro-F1 and macro-F1 measures of Naive Bayes baseline classifiers. As we explained earlier, one reason is that SVM has the ability to generate well in dimensional feature spaces, since SVM classifier exhibits overfitting protection that does not depend on the number of features. Generally the performance improvement of SVM classifier using feature selection method hasn't been proved to be significant. The fact that SVM classification result is improved by using positional information shows that the positional information is very important factor and can provide useful information to indicate what term is important. Also to improve the performance of SVM classifier, the feature

selection or term weighting methods need to determine the relevant features very accurately. By combining different factors in term weighting, we have been able to improve the performance of SVM classifier. This shows that although SVM itself has the ability to retain important features, some preprocessing can still be useful to build a better document representation of document, and to improve the performance of SVM.

As we can observe, generally SVM has better classification performance than NB. However, it is also more expensive in terms of computation resources. NB is much more efficient. In these experiments, we showed that by integrating additional features into document representation, we could obtain a classification performance close to that of SVM. This result is very encouraging and few previous experiments have produced a classification result with NB comparable to that of SVM. Our result shows that NB can be a very effective classifier, provided that appropriate features are used. In some domain-specific applications such as ours, well-tuned NB can be used as a more efficient replacement of SVM.

Our features vs. InfoGain

For Naive Bayes classifier, the experiments done with term weighting methods give better classification performance compared to the InfoGain feature selection method. The sentence filtering method shows similar classification performance compared to the InfoGain. For SVM classifier, only the term weighting method with all the factors combined gives performance slightly better than the InfoGain feature selection method. The reason might be that the Naive Bayes method is very dependent on good feature set. Methods with NE, concept and location information can eliminate document noise better

than the InfoGain method. In contrary, the SVM classifier is not very dependent on feature selection method. So feature selection methods do not help improve SVM.

CHAPTER 6

CONCLUSION AND RECOMMENDATION

Text classification has often been studied extensively for general texts. Little attention has been paid to the particular characteristics of documents to be classified. In this study, we argue that for a specific type of document, special processing can be useful to help increase the classification accuracy. In our study, we investigate the problem of classification of call for tender (CFT) documents. Unlike the typical news documents, CFT documents contain lots of procedural information unrelated to the subject of the documents, while relevant information is described only by a few sentences. Our hypothesis in this study is that classification results can be improved if we can select or weight features according to the particular characteristics of the documents.

In this dissertation, we examined three characteristics of the CFT documents: the position of the important sentences, the inclusion of different types of named entities in sentences and the inclusion of NAICS concepts in sentences. These three types of factors are used to select relevant features or to weight them.

We examined two ways to incorporate the three factors. One is feature selection by filtering sentences that are considered to be irrelevant to the subject. Another one is term weighting that incorporates additional factors of sentences.

To verify the effect of both proposed methods, we conducted experiments using Naïve Bayes and SVM classifiers on a set of CFT documents collected from Federal Business Opportunities website. These documents have been manually classified into NAICS classes. We observed the following facts in our experiments:

- With the sentence filtering method on Naïve Bayes classifier, we observed no significant improvement with named entities and location information compared to the baseline experiment. We obtained slight improvements with sentence filtering on concept, which is similar to the result of experiments with feature selection method.

- With the term weighting method on Naïve Bayes classifier, the results are generally superior to the result with the sentence filtering method. We got an increase of 8% in micro-F1 over the baseline NB on the term weighting with concept, an increase of 10% of micro-F1 on the term weighting with named entities and an increase of 10% of micro-F1 on the term weighting with position information. With all the factors combined together, we obtained an increase of 15% of micro-F1.

- For the SVM classifier, we didn't get any classification result better than the baseline with the feature selection and term weighting method on named entities and concepts used separately. However, with the term weighting method that used all three factors and term weighting method with position information, the classification result is slightly better than the baseline.

- For this particular collection, we found that both feature selection by sentence filtering and term weighting method improves the performance of Naive Bayes classifier by a bigger margin than the performance of SVM classifier. Also term

weighting methods give better classification result than the sentence filtering methods. Among all the factors studied, NE and location information seem to be the factors that can best help to distinguish the relevance and irrelevant parts of documents and give better performance than the concept factor. The method that combines all the factors together give the best performance for both SVM and Naive Bayes classifier.

- Compared to the standard feature selection method like InfoGain, for Naive Bayes classifier, sentence filtering method and term weighting method by all three factors give better performance. However, for SVM classifier, only the term weighting with position information and term weighting methods that combines three factors give better performance than InfoGain feature selection method.

As a general conclusion, we can say that it is useful to integrate specific characteristics of the documents into document representation before using general classification methods. This is particularly important for NB classifier.

In this study, we have examined the possible impact of three additional factors on classification. However, we have not tried to determine the best way to exploit them. This will be an interesting future research topic.

The proposed approach of term weighting in this thesis can also be useful for information retrieval (IR). For example, we can assign higher weights to the features in the first few sentences of the document in the indexing process of the information retrieval. However, the lack of an appropriate test collection poses a problem for us to test this approach.

REFERENCES

Brill, E. (1993), *A Corpus-Based Approach to Language Learning* , University of Pennsylvania, Department of Computer and Information Science, Ph.D. Thesis, IRCS Report 93-44, 166 p.

Cavnar, W. (1993) N-gram-based text filtering for trec-2. In: *Second Text Retrieval Conference (TREC)*. pp 177-178.

"concept." Merriam-Webster Online Dictionary. 2007. http://www.merriam-webster.com (5 Jan. 2007).

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., and Ursu, C. (2002) The GATE User Guide. http://gate.ac.uk/.

Edmundson, H. P. (1969) New methods in automatic abstracting. *Journal of the ACM*, 16(2): 264–285.

Gospodnetic, R. and Hatcher, E. (2005). *Lucene In Action.* Manning Publications.

Gaizauskas, R., Wakao, T., Humphreys, K., Cunningham, H., Wilks, Y. (1995) Description of the LaSIE system as used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. Morgan Kaufmann, California. pp 11.

Joachims, T., (1998) Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Proceedings of ECML-98, 10th European Conference on Machine Learning*. Heidelberg, Germany, pp 7-8.

Ko, Y., Park, J., Seo, J. (2002) Automatic Text Categorization using the Importance of Sentences *International Conference On Computational Linguistics Proceedings of the 19th international conference on Computational linguistics - Volume I.* pp 1-3.

Kupiec, J., Pedersen J., Chen, F., (1995) A Trainable Document Summarizer, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp 68-73.

Lemay, M. (2006) NConcept Extractor. http://uima.lti.cs.cmu.edu:8080/UCR/pages/static/nstein_NConcept_extractor.htm (5 Jan. 2007).

Lin, C. Y. (1995) Knowledge-based automatic topic identification. *Proceedings of the Association for Computational Linguistics (ACL-95)*, pp 308-310.

McCallum, A. K. (1996) Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/~mccallum/bow. (5 Jan. 2007).

McCulloch, D. (2004) An Investigation into Novelty Detection http://www.enm.bris.ac.uk/teaching/projects/2004_05/dm1654/svm_classification.htm. (5 Jan. 2007).

Mitchell, T. (1996) Machine Learning. McGraw Hill.

Mock, K.J. (1996) Hybrid hill-climbing and knowledge-based techniques for intelligent news filtering. *Proceedings of the National Conference on Artificial Intelligence (AAAI'96)*. Menlo Park, California, pp 8-10.

"NER." Wikipedia. 2007. http://en.wikipedia.org/wiki/Named_entity_recognition (5 Jan. 2007).

Quinlan, J. R. (1993) *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Salton, G., Yang, C., and Wang, A. (1975) A vector space model for automatic indexing. *Communications of the ACM*, 18(11), pp 613-620.

Salton, G., Fox, E.A., Wu, H. (1983) Extended Boolean information retrieval. *Communications of the ACM,* 26 (12), pp 1022-1036.

Vapnik, V. (1995) The Nature of Statistical Learning Theory. Springer, New York.

Yang Y., J.O.P. (1997) A comparative study on feature selection in text categorization. In: Proceedings of 14th International Conference on Machine Learning, Nashville, US, pp 2-4.

Yang, Y. (1999) An evaluation of statistical approaches to text categorization. *Inf. Retr.*, 1(1-2), Carnegie Mellon University, pp 69–90.